AFHRL-TP-89-60

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

AD-A220 821

GENERALIZABILITY OF PERFORMANCE MEASURES ACROSS FOUR AIR FORCE SPECIALTIES

DTIC
ELECTE
APR 24 1990
S D
B

Kurt Kraiger

University of Colorado at Denver
Department of Psychology
1200 Larimer Street
Denver, Colorado 80204

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

April 1990
Interim Technical Paper for Period October 1987 - September 1989

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

90 04 19 031

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

MARK TEACHOUT
Contract Monitor

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

RODGER D. BALLENTINE, Colonel, USAF
Chief, Training Systems Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1990	Interim - October 1987 to September 1989

4. TITLE AND SUBTITLE
Generalizability of Performance Measures Across Four Air Force Specialties

5. FUNDING NUMBERS
C - F41689-86-D-0052
PE - 62205F
PR - 1121
TA - 13
WU - 01

6. AUTHOR(S)
Kurt Kraiger

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Colorado at Denver
Department of Psychology
1200 Larimer Street
Denver, Colorado 80204

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Training Systems Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFHRL-TP-89-60

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Generalizability (G) theory was used to assess the reliability of criterion measures across four Air Force specialties. This analysis strategy was applied to work sample tests (i.e., Walk-Through Performance Test (WTPT) scores) and ratings of job proficiency. Results indicated that WTPT scores were reliable within each specialty. In addition, ratings were reliable across different rating forms (i.e., general to specific) and within rating sources (i.e., incumbents, peers, and supervisors). However, ratings were not generalizable across rating sources, and ratings were not substitutable for WTPT scores. These results were consistent across the four specialties, lending increased confidence to the findings.

14. SUBJECT TERMS
G-theory, generalizability, job performance measurement, performance ratings, reliability, walk-through performance test, work sample test

15. NUMBER OF PAGES
108

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFHRL Technical Paper 89-60

April 1990

GENERALIZABILITY OF PERFORMANCE MEASURES ACROSS FOUR AIR FORCE SPECIALTIES

Kurt Kraiger

University of Colorado at Denver
Department of Psychology
1200 Larimer Street
Denver, Colorado 80204

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

James B. Bushman, Major, USAF
Chief, Training Assessment Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

This paper presents an application of generalizability (G) theory to the Air Force Job Performance Measurement (JPM) Project. G theory is briefly reviewed, then applied as a data analysis strategy for proficiency ratings and Walk-Through Performance Test (WTPT) scores in four occupational specialties. The primary findings were that WTPT scores were dependable within each specialty, as were proficiency ratings within rater sources (i.e., incumbents, peers, and supervisors). Ratings were not generalizable across rater sources, and ratings were not related to WTPT scores. The similarity of results across the four specialties lends increased confidence to the findings. The results also support assertions that the WTPT is the high fidelity measure in the JPM project.

Accession For	
NTIS GRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

PREFACE

The Air Force Job Performance Measurement (JPM) project is a remarkably broad, encompassing attempt to assess individual job proficiency. Within the four specialties examined here, incumbents are assessed via a Walk-Through Performance Test (WTPT), with hands-on and interview components, and subjective job proficiency ratings. The ratings themselves are broad-based, as incumbents are evaluated, on four different forms, by themselves, their peers, and supervisors.

A critical issue concerns the psychometric quality of these various measures. The present study addresses this issue by: assessing the psychometric quality of both the WTPT and rating methods; examining the extent to which the ratings are substitutable for the WTPT; and comparing the results across specialties. The psychometric quality of the proficiency ratings and the WTPTs (as well as their substitutability) was assessed via generalizability (G) theory. G theory identifies whether scores assigned to individuals are dependable (or consistent) over conditions of measurement. For the rating data, the relevant conditions of measurement were rater sources, rating forms, and items or dimensions within particular forms. For the WTPT data, relevant conditions of interest were assessment method (hands-on or interview), tasks, and steps or items within tasks. For the question of substitutability, a third generalizability design was constructed with performance

measures (ratings or WTPT scores) and tasks as the conditions of interest. In addition to assessing the generalizability of these measures under current assessment conditions, generalizability theory was used to forcast the expected dependability of these measures under reduced measurement conditions (e.g., a single rating source or a single WTPT method).

The author is grateful to Mr. Mark Teachout of the Training Systems Division, Air Force Human Resources Laboratory, for his assistance in the completion of this paper. Mark aided the completion of this paper by sharing his knowledge of the JPM project and by his timely review of earlier drafts of this manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE Page

GENERALIZABILITY OF PERFORMANCE MEASURES ACROSS FOUR AIR FORCE SPECIALTIES

I. INTRODUCTION

The major goal of the Air Force Job Performance Measurement (JPM) project is to provide the data necessary to establish valid linkages between enlistment standards and job performance. Such an effort depends on the development of empirically sound measures of performance. To date, the project staff has developed both Walk-Through Performance Tests (WTPT) and proficiency rating measures and applied them to data collection in four specialties. The present research supports the development of these measures by: assessing the psychometric quality of both the WTPT and the performance ratings, examining the extent to which the proficiency ratings are substitutable for the WTPTs, and comparing these results across the four specialties to identify appropriate measurement conditions for future efforts.

Generalizability (G) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is applied to the first two research issues. In G theory terms, the primary issues are whether each evaluation system yields dependable scores over conditions of measurement, whether incumbent performance levels are dependable over various evaluation methods, and whether conclusions drawn from the first two investigations are similar over occupational specialties.

This introduction will be organized as follows. First, the various performance measurement methods will be briefly described. Next, G theory will be formally introduced and described. Finally, results will be presented and interpreted.

Current Job Performance Measures

The current Air Force JPM uses both work sample measures and performance ratings to assess incumbent proficiency. More complete details of the JPM system are given in Hedge and Teachout (1986).

The benchmark method, known as Walk-Through Performance Testing (WTPT), includes both observation of actual hands-on performance, and incumbent interview testing. The WTPT hands-on format requires job incumbents to perform a series of job tasks under the careful observation of a highly-trained test administrator. The interview format requires incumbents to describe in detail the steps they would perform to accomplish various job tasks. Since the WTPT is performed at the work site, the incumbent is able to visually refer to necessary equipment, tools, work supplies, etc. With both formats, administrators record whether or not critical behaviors were performed (or described).

In addition to the WTPT, incumbents are assessed on four different rating forms by three different sources: Incumbents themselves, one to three peers, and immediate supervisors. Each rating form assesses individual

proficiency via a 5-point rating scale, with specific behavioral descriptions provided for scalar points on three of the forms.

The most specific rating data are provided by the task rating form which solicits proficiency ratings on a representative sample of tasks. The second most specific data are provided by the dimensional rating form which solicits proficiency ratings on broad groupings of tasks, identified through factor analysis of occupational survey data and input from subject-matter experts. The third most specific form is the global rating form which includes only two items covering the job domain: (a) technical proficiency and (b) interpersonal proficiency. The most general rating form is the Air Force-wide form, which includes general performance factors developed to be representative of all specialties in the Air Force.

Overview of Generalizability Theory

Generalizability theory was developed by Cronbach and his associates (Cronbach et al., 1972; Cronbach, Rajaratnam, & Gleser, 1963) as an alternative to classical test theory. Whereas classical theory permits only univariate investigations of the effects of measurement error on reliability, G theory permits multifaceted analysis of the dependability of scores over a variety of measurement conditions. Recent detailed discussions and reviews of generalizability theory may be found in Brennan (1983), Brennan and Kane (1979), Kraiger (1989), and

Shavelson, Webb, and Rowley (1989). The rudiments of generalizability theory are explained below.

Generalizability theory answers the question, "Does it matter if . . . ?" That is, generalizability analyses can determine the relative variance in scores which can be attributable to various conditions of measurement. If variance over conditions is low, overall scores are said to generalize over the conditions of measurement. More informally, low variability over conditions implies that it "doesn't matter if" the measure is operationalized in different ways. Said yet another way, generalizability analyses indicate the degree to which scores based on a limited opportunity for observation (e.g., a work sample on a single occasion) are dependable over a considerably broader sample of possible observations (e.g., other tasks, occasions, etc.)

In any generalizability study, the researcher must first identify any factors of interest which could affect the measurement process. The researcher then must specify a particular range of levels for each factor. In G theory terminology, factors of measurement are called facets and levels of the facet are called conditions. An individual's average score over all combinations of conditions is said to be that person's universe score. Thus, if a researcher is conducting a generalizability analysis on performance ratings, he or she may specify raters and occasions as facets of interest. That is, the researcher wishes to know

whether performance ratings vary substantially over judges or occasions. The actual number of raters within the organization or the intended number of rating occasions may determine the number of levels for the facets.

A generalizability study (G) study could be designed to estimate the contribution to total score variance of the following sources: Individual performance levels, raters, occasions, and all interactions involving these three sources. Data for the generalizability study would come from random samples of ratees, raters, and occasions. Variance components would be estimated for each source. Variance components represent estimated variance about universe scores for average single observations, e.g., an average person evaluated on an average rater on a single occasion. In addition, a summary generalizability coefficient could be computed from individual variance components. This coefficient is analogous to a reliability coefficient in classical test theory and represents the proportion of observed score variance which is attributable to individual differences in the attribute being assessed. However, interpretations of individual variance components are often more enlightening since these reflect contributions to error variance by particular aspects of the measurement system (Brennan & Kane, 1979) and may be interpreted as evidence of construct validity (Kraiger & Teachout, 1989).

While G study analyses are useful for identifying the general characteristics of a measuring device, the same analyses may be misleading for describing the psychometric quality of an instrument under actual or intended circumstances. This is because G study variance components are estimated for single items or single administrations, even though organizations typically use multiple operationalizations of constructs (e.g., multiple-item scales). Thus, a researcher may wish to perform a decision (D) study to assess the specific characteristics of a measurement instrument in a particular decision-making context. Similar to the Spearman-Brown prophecy formula in classical test theory, D studies allow a researcher to forecast resulting variance components and generalizability coefficients under different sets of measurement conditions. While the Spearman-Brown formula permits estimation when only a single parameter (typically items) is varied, D studies allow estimation of estimated effects when multiple facets are simultaneously varied. For example, generalizability coefficients can be estimated when ratings are averaged over three raters on a single occasion or two raters on two occasions. It is often these D study results which are of the most interest to decision-makers since they reflect realistic or intended measurement conditions.

Applications of Generalizability Theory in the Military

For several years, various branches of the military have shown a marked interest in applications of generalizability theory to their respective performance measurement projects. Both Kraiger (1989) and Shavelson (1986) reviewed generalizability theory and discussed its relevance to job performance measurement research. Shavelson concluded that generalizability theory was "the most appropriate behavioral measurement theory for treating military measures" (p. 61) because it models the sources of error likely to enter into a performance measure, provides information on where the major sources of measurement error lie, provides estimates of improvement in measurement under alternative sampling plans, and suggests alternative revisions in measurement strategy when sampling alone is insufficient for overcoming measurement error.

At a recent conference of the American Educational Research Association, representatives of the Navy, Army, and Air Force reviewed their recent work in the area of generalizability theory. Webb and Shavelson (1987) assessed the generalizability of hands-on performance tests for 27 Machinist Mates. For the hands-on test, facets of interest were test examiners and tasks composing the test. Tasks were found to be a major source of error variance as Machinist Mates were rank-ordered differently by tasks. Examiners were not a major source of error variance. Each examiner gave similar mean scores across ratees and also

rank-ordered ratees similarly. Webb and Shavelson concluded that sufficient levels of generalizability could be achieved by using only a single rater, but only if a considerable number (18) were assessed on the hands-on test.

In the same session, McHenry, Hoffman, and White (1987) presented a generalizability analysis of performance ratings for 7,045 soldiers in 19 Army jobs. For their analyses, rater type (peers and supervisors) and rating scale were the facets of interest. Analyses were performed for each job and within each of three general performance factors identified in previous analyses by McHenry et al. When scores were averaged over rater type, generalizability coefficients were very high for two of the three performance factors (effort/leadership and personal discipline, but not physical fitness/military bearing). Inspection of the individual variance components revealed that generalizability was lower on the third factor because of a large interaction between rating scales and ratees. In other words, ratees were differentially ranked across scales comprising the physical fitness/military bearing factor.

Finally, Kraiger and Teachout (1987, 1990) presented analyses conducted on 256 Air Force jet engine mechanics. Facets of interest were rater type (self, peer, and supervisor), rating forms (four different forms were used), and items or scales nested within forms. Results were

interpreted as evidence of construct validity, with specific variance components revealing fairly high convergence over forms but not over rater sources. There was also little variance due to items. Kraiger and Teachout concluded that there was moderately strong evidence for the construct validity of the ratings and that generalizability theory was a useful tool for analyzing and understanding rating data.

As preliminary applications of generalizability theory to performance measurement data have been successful, it appears useful to continue using the G-theory for other additional analyses. In particular, generalizability analyses can also be applied to WTPT scores, or to designs which include ratings and WTPT scores. Another question of interest is whether generalizability estimates are consistent across other applications. S e the Air Force Performance Measurement Project has continued data collection in three other specialties, the decision was made to extend the G-theory methodology to these specialties as well. The principal research questions for this investigation are detailed below.

Current G-Theory Investigation

Overview. To date, complete performance measures have been developed for and data collected in four Air Force specialties. For each specialty, performance data included performance ratings on task-level, dimensional, global, and Air Force-wide rating form from incumbents, peers, and

supervisors; WTPT hands-on performance measures; and WTPT interview measures.

Generalizability of Rating Data. The first area of inquiry concerns the generalizability of performance ratings over different conditions of measurement. This investigation seeks to replicate the findings of Kraiger (1989; Kraiger & Teachout, 1987, 1989) in the other three specialties. As noted below, the facets of interests are rating sources (incumbents, peers, and supervisors), forms, and items nested within forms. Results both within and across specialties are of interest.

Generalizability of WTPT Data. The second area of interest addresses the generalizability of the WTPT scores. For each specialty, incumbents are assessed using both the hands-on and interview formats. Facets of interest are specific methods (hands-on and walk-through), the number of tasks assessed by either method, and the number of items or steps comprising individual tasks. Results both within and across specialties are of interest.

Substitutability Design. The extent to which performance ratings can be considered acceptable surrogates for the more extensive WTPTs is another important issue which can be addressed through generalizability analyses. In this design, methods (ratings, hands-on, and interview) and tasks were considered the primary facets. Separate analyses were performed for all rating sources combined, as

well as each source individually. Again, results are to be examined both within and across specialties.

II. METHOD

Sample

Proficiency ratings were collected from first-term airmen in four different specialties. The Air Force specialties (AFS) and their respective sample sizes were: Jet Engine Mechanic (AFS426x2), n = 255; Air Traffic Control Operator (AFS272x0), n = 172; Avionic Communications Specialist (AFS328x0), n = 98; and Information Systems Radio Operator (AFS492x0), n = 156. The generalizability analyses described below were repeated in each specialty.

Rating Facets of Generalization

For the investigations of the performance rating data, there were three facets of generalization: Rating forms, specific items or scales included on each form, and rating sources. Items were nested with forms, and both were crossed with sources and ratees. As illustrated in Figure 1, there were 11 distinct sources of variance: Persons (or ratees), sources, forms, items within forms, forms by sources, persons by sources, persons by forms, persons by items within forms, sources by items within forms, persons by forms by sources, and persons by sources by items within forms. In addition, random error, σ^2_e, is confounded in the design with the latter term. Variance due to persons was considered desirable (since it results from individual

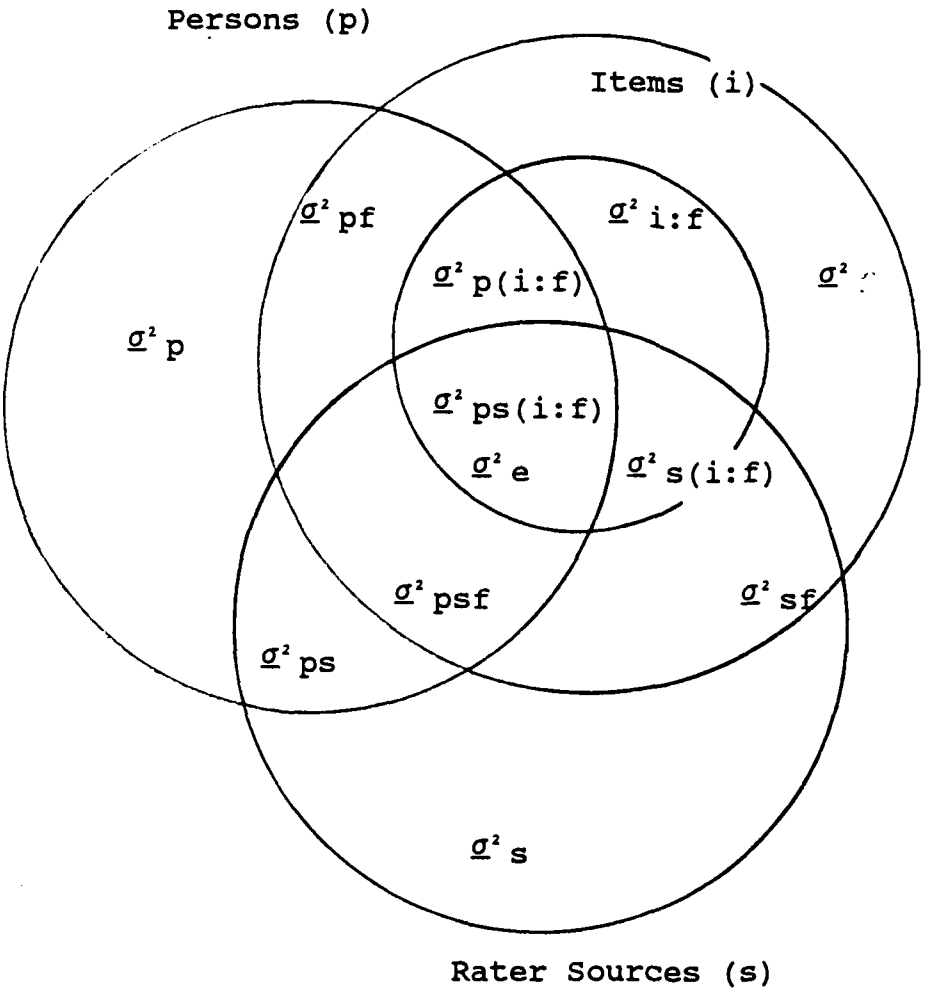


Figure 1. Venn Diagram Illustrating Variance Components for Rating Variables Generalizability Design.

differences among ratees in performance), while all other sources contribute to absolute or relative error variance. Relative error variance affects comparative or ranking decisions and is comprised of all variance components which represent an interaction of the person facet and at least one other facet (e.g., σ^2_{pf}). Absolute error variance

affects decisions about individuals in reference to an absolute cutoff and contains all variance components in Figure 1 except σ^2_p. Relative error variance is of greater interest when criterion scores are used in validation designs, while absolute error variance would be used if criterion scores were also used for selection or minimum competency decisions.

Complete details concerning the treatment of each facet are given in Kraiger (1989). The first facet of interest was rating sources. In order to balance the design (see below), when airmen were rated by more than one peer, only a single randomly-selected rating was used. Low estimates for variance components involving sources would indicate that the use of different sources provides little error variance to the measurement process. A second facet was rating forms, since four different rating forms (task-level, dimensional, global, and Air Force-wide) were used for data collection. The actual forms used can be considered random samples of a larger universe of possible forms which could be used to assess ratee performance. Low estimates for variance components including the forms facet would indicate that little error variance is introduced by the use of different rating methods. The final facet was the individual items, dimensions, or scales which comprise each form. Again, the items on any one form can be considered a random sample of possible items which could constitute that form. As seen in Figure 1, items are

nested within forms because individual items or scales vary from form to form.

As noted in Kraiger (1989), there is a computational problem with this facet since the number of items on a form can range from two (on the global form) to over 30 (on the task-level form). In standard analysis of variance (ANOVA) terms, this means that the items facet is unbalanced since there is a different number of conditions under each of the forms conditions. Searle (1971) has shown that unbalanced ANOVA designs produce biased mean square estimates, which in turn are used to compute variance components. In general, experts in generalizability theory recommend against the use of unbalanced design (Brennan & Kane, 1979; Shavelson & Webb, 1981).

As in Kraiger (1989), two strategies were used to create balanced items across forms. The first involved randomly selecting only two items from all four forms. While this strategy allows analyses across all forms, considerable information about variability in items may be lost. The second strategy was to exclude the two-item global form and analyze data from the other three forms with a greater number of items or scales. In all specialties, the next smallest number of items was on the dimensional form. Thus, the number of scales on this form determined the number of items randomly selected for the other forms. For example, for Air Traffic Control Operators, there were only five scales on the dimensional

form, so five items were randomly selected from the Air Force-wide and task-level forms. Analyses for both the four form and three form designs are presented and compared in the results section.

Facets for WTPT Data

For G study investigations of the WTPT data, there were three facets of interest. The first was the method of assessment, hands-on vs. interview. In the hands-on component, the incumbent actually performs the test task under the careful observation of the test administrator. In the interview component, the incumbent describes the steps he or she would take to complete the task. Since all WTPTs are performed at the work site, the incumbent is able to refer to actual tools, manuals, and equipment during the testing. Low estimate of variance components for the method facet would indicate that varying the method of assessment adds little error to the WTPT measurement process.

The second facet of interest was the tasks that were measured by both the hands-on and interview components. Typically, a WTPT consisted of about 20 tasks. For each specialty, these tasks can be considered random samples of a larger possible universe of tasks which could comprise the WTPT. Low estimates of the variance component for the task facet would indicate that little error variance is introduced to the measurement process through the sampling of tasks. Also, low variance component estimates would

suggest that fewer tasks could be used in future applications without appreciable losses in measurement fidelity.

There were two different possible designs for analysis of the generalizability of scores over tasks. For each specialty, there were three types of tasks included in the WTPT: Tasks common to both the hands-on and interview components, tasks unique to the hands-on component, and tasks unique to the interview component. Consider the common tasks. It is clear that these tasks should be considered crossed with method, since each task is assessed by each method and each method includes all tasks. This design is illustrated by the Venn diagram in Figure 2. Unique tasks, however, should be considered nested within methods. This design is illustrated in Figure 3. The nesting indicates that a task assessed by one method was not measured by the other method. To increase the number of task conditions analyzed (and reduce sampling error in the entire design), analyses were conducted with the common tasks considered nested along with the unique tasks. That is, eight unique tasks and six common tasks might be nested within a method, even though these common tasks were not really nested. Results of these analyses were very similar to results from analyses using only unique tasks, but with smaller sampling error in the estimates of variance components. Results of the larger design with both unique and common tasks treated as nested are presented below.

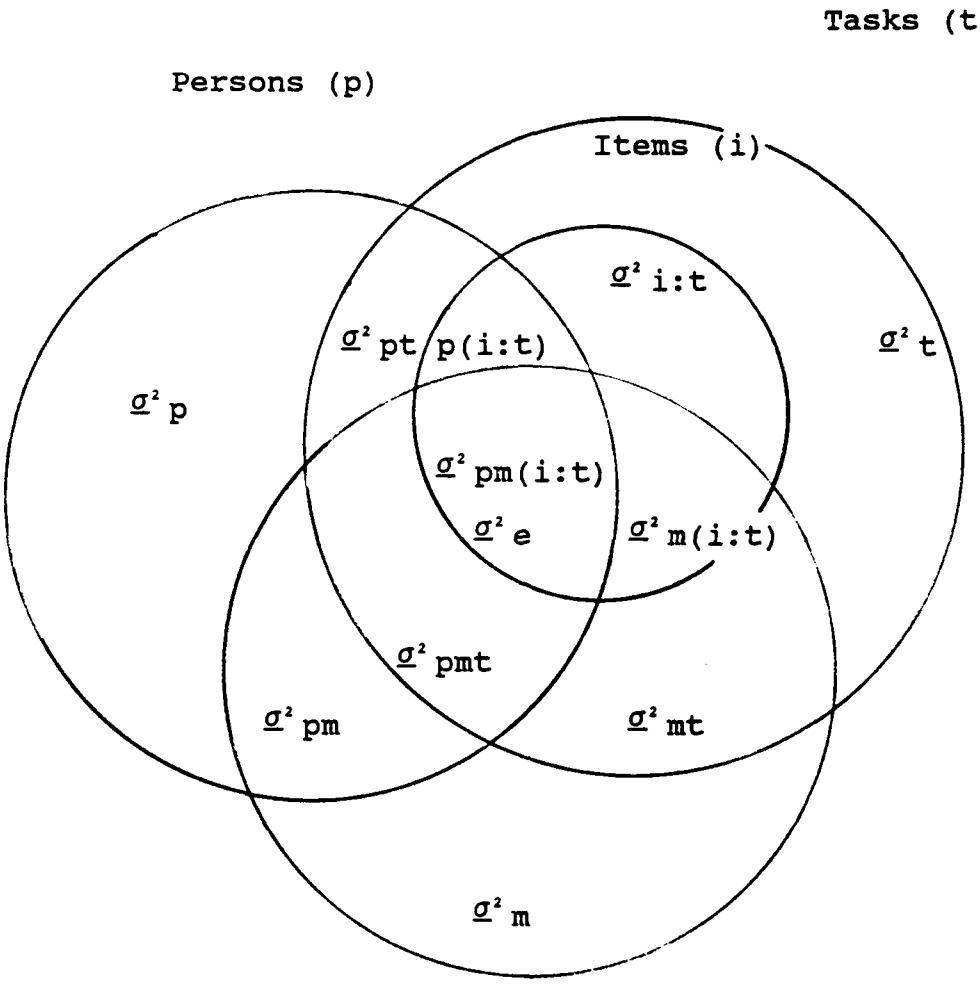


Figure 2. Venn Diagram Illustrating Variance Components for Generalizability Analysis of WTPT Data with Tasks Crossed with Methods.

For some specialties, there were uneven numbers of tasks across the two methods. To balance the design, one or two unique tasks were randomly omitted from analyses.

The final facet of interest was the number of items or steps comprising individual tasks on the WTPT. Items were

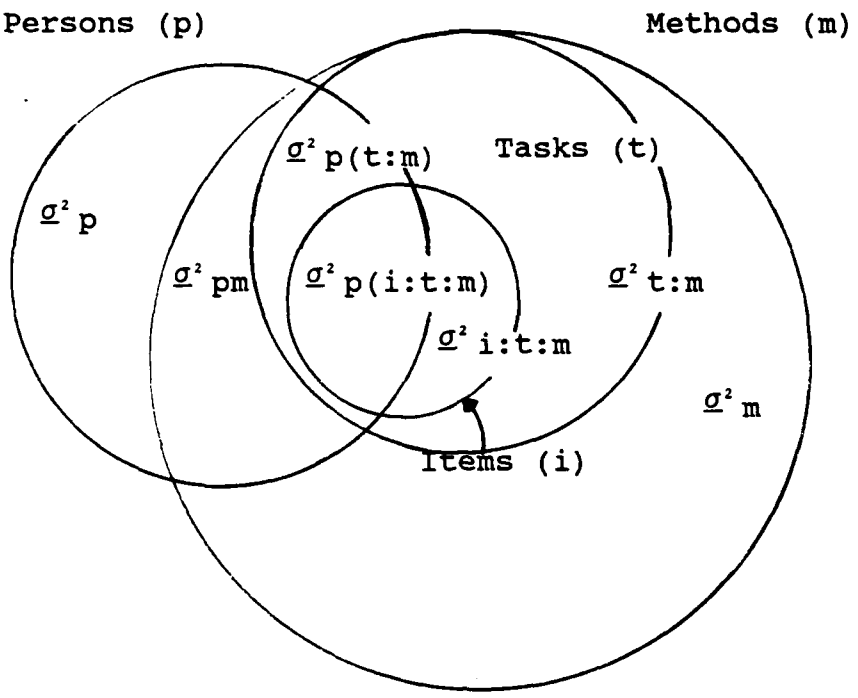


Figure 3. Venn Diagram Illustrating Variance Components for Generalizability Analysis of WTPT Data with Tasks Nested in Methods.

treated as nested within tasks since they were in fact different for each task of the WTPT. For each task, the items can be considered random samples of larger possible universes of possible items. Low estimates of variance components involving the items facet would indicate that incumbents scores do not vary over individual items. Perhaps more importantly, low estimates would suggest that fewer items could be used to assess task performance in later applications of the WTPT without appreciable losses in measurement fidelity.

As in the case of the items facet for the rating analyses, the items facet for the WTPT was unbalanced since the number of steps for a task ranged from as little as four to greater than 30. Depending on the specialty, tasks with as few as three, four, or five items were excluded from the analysis, and the next smallest number of items on a task was used as the number of conditions for the items within tasks facet. That many items were randomly selected from all other tasks included in the design. For example, for the Information Systems Radio Operator, tasks with less than six items were not analyzed, and six items were randomly sampled from all tasks with more than six steps.

Facets for Substitutability Design

The final generalizability design was used to assess the extent to which the assessment of individuals' proficiency levels were generalizable over the three primary measurement methods: Ratings, hands-on testing, and interview testing. Thus, the main facet of interest was evaluation method. Low estimates for variance components involving the methods facet would indicate that individuals' proficiency scores were consistent over methods of assessment.

For the ratings condition of the method facet, only task-level ratings were analyzed. Since this form was designed to have the greatest overlap in content to the WTPT, it was assumed that generalizability coefficients using the task form would be greater than those generated

from analyses of any other form. A second issue concerned the appropriate rating source for the analyses. It was decided to address this issue empirically rather than theoretically so that four different analyses were conducted -- one with ratings of each source compared to the walk-through scores and a fourth with ratings averaged across the three sources.

The other facet of interest was the number of tasks which could be measured by each method. For purposes of the G study analyses, this was defined by the smaller number of tasks which constituted either the hands-on or interview component of the WTPT for an AFS. An equivalent number of tasks were randomly sampled from the other WTPT component and from the task-level rating form.

Tasks and methods were considered crossed, that is, it was assumed that all tasks were measured by each method, and each method assessed each task. A Venn diagram illustrating the substitutability design is shown in Figure 4.

D Study Analyses

After all G studies were completed, the final stage of analyses was to present simulated D study analyses. Recall that G study variance components represent the estimated variability about universe scores for average single observations, e.g., one person evaluated on an average task by a typical administrator. Such analyses are primarily useful for examining the relative contributions of various

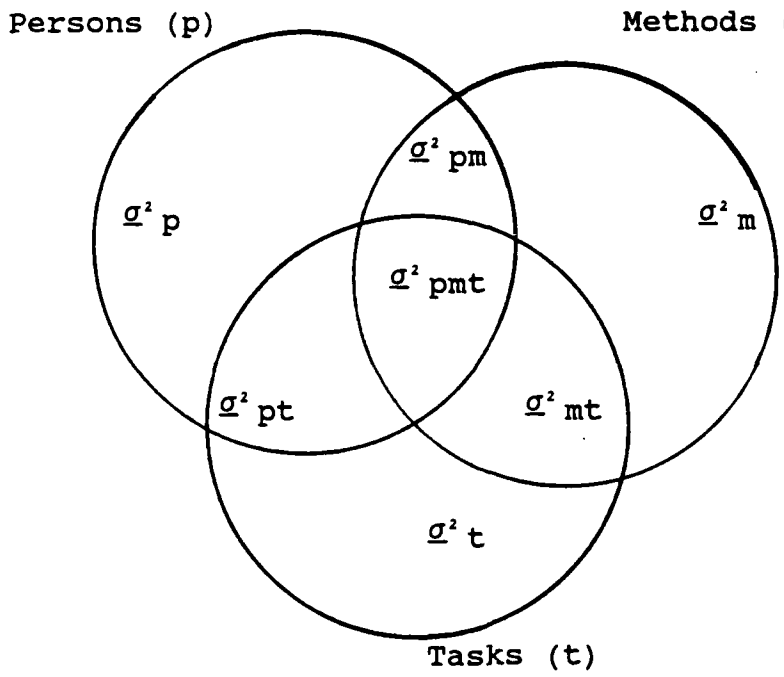


Figure 4. Venn Diagram Illustrating the Substitutability Design for Analyses of Ratings, Hands-on Scores, and Interview Scores.

sources of variance. These analyses are misleading though when organizations employ multiple dimensions, occasions, or raters with their measures (e.g., more than one dimension). Not unexpectedly, G theory is able to account for the reduction of measurement error when individuals' scores are averaged over multiple items, scales, tasks, etc. The purpose of D study analyses is to assess the effects of these multiple operationalizations of instruments on measurement error.

Although new data may be collected for D study analysis, the most typical input is previously generated G

study results. When random facets are assumed, typical D study results include adjusted variance components for individual effects, total universe score variance (variance due to individual differences, often $\underline{\sigma}^2_p$), relative error variance ($\underline{\sigma}^2_\delta$, equal to the sum of all effects which contain p and at least one other index), absolute error variance ($\underline{\sigma}^2$, equal to the sum of all effects in the design except $\underline{\sigma}^2_p$), and their associated generalizability coefficients ($\underline{\epsilon P}^2$, for relative decisions; and $\underline{\Phi}$, for absolute decisions).

Conditions in the D study are defined by how the organization uses (or intends to use) the instrument. A complete discussion of the delineation of D study parameters is found in Gillmore (1979; 1983). For present purposes, only two D study parameters will be discussed: The treatment of a facet as fixed or random and the number of conditions within the facet.

Fixed and Random Facets. Random facets imply that the conditions of a facet represent a random sample from an essentially larger set of possible cases. At a minimum, the researcher must be willing to assume that the conditions sampled in the study could be replaced with other elements of some larger set of possible observations without affecting the universe score (Shavelson & Webb, 1981). When a random facet is specified, generalization is not limited to the set of D study conditions, but instead extends to the entire range of admissable observations. In

contrast, a fixed facet implies that the conditions observed in the G study exhaust the range of possible conditions of interest to the organization and that the organization intends to use an average or total score over conditions of the facet. While the distinction between fixed and random facets is meaningless at the G study level (all facets are treated as random), it is important at the D study level in that the specific variance components which enter into computations of universe and error variance. For the three designs analyzed in this study, an assumption of a fixed facet was made only for the WTPT methods. D study results for that design were analyzed with methods (hands-on and interview) as both a random and fixed facet.

Number of Conditions. A second decision permitted at the D study level concerns the number of conditions observed for each facet. The error variance attributable to any facet (and its interactions) is reduced as scores are summed or averaged over greater numbers of conditions of that source. Likewise, generalizability coefficients increase as conditions are increased and error variance decreased. Just as the Spearman-Brown prophecy formula allows estimation of a test's univariate reliability as test length is increased, D studies permit the estimation of changes in variance components and generalizability coefficients as the conditions of multiple facets are increased or decreased. Thus, for each design above,

multiple combinations of D study conditions were proposed (e.g., 10 items on 10 tasks with one WTPT method or 15 items on 5 tasks with both methods) and the resulting adjusted variance components and generalizability coefficients were forecast. Operationally, a D study variance component is adjusted by dividing the variance component by the number of conditions of any facet indicated by its subscript. For example, the G study estimate for $\underline{\sigma}^2_{i:f}$ would be divided by 30 if 10 items on each of three forms was specified as a set of D study conditions.

The notation used for D studies should be explained. Consistent with Brennan's recommendations (Brennan, 1983; Brennan & Kane, 1979), D study facets are noted by capital letters in the subscript. The "p" associated with individuals remains lower-case since persons are not treated as a facet in these analyses. Thus, the G study effect $\underline{\sigma}^2_{i:f}$ is indicated as $\underline{\sigma}^2_{I:F}$ at the D study level, while $\underline{\sigma}^2_{ps}$ is indicated as $\underline{\sigma}^2_{pS}$.

III. RESULTS

Ratings Design

Descriptive Results. Tables 1 through 4 present traditional analyses of the rating data for each specialty. Presented within combinations of rating form and rating source are the average dimension/item mean and the average dimension/item intercorrelation. Also presented in the tables are averaged correlations indicating convergent

Table 1. Descriptive Statistics for Rating Variables, with Jet Engine Mechanics

Source: Form	$\underline{m}^a$	$\underline{r}^b$	$\underline{r}^a$ with Self	Sup.[c]	Peer
Self					
Task	4.02	.30	--	.11	.15
Dimensional	3.80	.41	--	.31	.34
Global	4.13	.38	--	.28	.22
Air Force	3.74	.37	--	.27	.25
Supervisor					
Task	3.84	.53	.11	--	.13
Dimensional	3.55	.58	.31	--	.40
Global	3.86	.53	.28	--	.51
Air Force	3.51	.58	.27	--	.36
Peer					
Task	3.94	.49	.15	.13	--
Dimensional	3.66	.55	.34	.40	--
Global	3.80	.41	.22	.51	--
Air Force	3.45	.50	.25	.36	--

[a]averaged across dimensions within form

[b]average dimension intercorrelations within forms, sources

[c]Supervisor

Table 2. Descriptive Statistics for Rating Variables, with Avionic Communication Specialists

Source: Form	$\bar{m}^a$	$\bar{r}^b$	$\bar{r}^a$ with Self	Sup.[c]	Peer
Self					
Task	3.99	.60	--	.18	.25
Dimensional	4.03	.40	--	.37	.22
Global	4.04	.09	--	.31	.18
Air Force	3.79	.63	--	.24	.24
Supervisor					
Task	3.95	.51	.18	--	.26
Dimensional	3.89	.49	.37	--	.40
Global	3.83	.21	.31	--	.38
Air Force	3.63	.43	.24	--	.38
Peer					
Task	3.87	.42	.25	.26	--
Dimensional	3.95	.61	.22	.40	--
Global	3.86	.45	.18	.38	--
Air Force	3.59	.52	.24	.38	--

[a]averaged across dimensions within form

[b]average dimension intercorrelations within forms, sources

[c]Supervisor

Table 3. Descriptive Statistics for Rating Variables, with Air Traffic Control Operators

Source: Form	$\bar{m}^a$	$\bar{r}^b$	$\bar{r}^a$ with Self	$\bar{r}^a$ with Sup.[c]	$\bar{r}^a$ with Peer
Self:					
Task	4.04	.32	--	.22	.32
Dimensional	3.97	.41	--	.24	.25
Global	4.04	.46	--	.18	.21
Air Force	3.89	.39	--	.14	.15
Supervisor:					
Task	3.64	.45	.22	--	.26
Dimensional	3.60	.56	.24	--	.35
Global	3.69	.41	.18	--	.38
Air Force	3.52	.48	.14	--	.24
Peer:					
Task	3.88	.47	.32	.26	--
Dimensional	3.86	.49	.25	.35	--
Global	3.87	.51	.21	.38	--
Air Force	3.68	.43	.15	.24	--

[a]averaged across dimensions within form

[b]average dimension intercorrelations within forms, sources

[c]Supervisor

Table 4. Descriptive Statistics for Rating Variables, with Information Systems Radio Operators

Source: Form	$\underline{m}^a$	$\underline{r}^b$	$\underline{r}^a$ with Self	Sup.[c]	Peer
Self:					
Task	4.23	.44	--	.36	.35
Dimensional	4.22	.50	--	.28	.29
Global	4.24	.28	--	.25	.31
Air Force	4.03	.41	--	.24	.14
Supervisor:					
Task	4.29	.49	.36	--	.28
Dimensional	4.16	.51	.28	--	.30
Global	4.06	.37	.25	--	.39
Air Force	3.78	.48	.24	--	.23
Peer:					
Task	4.25	.38	.35	.28	--
Dimensional	4.17	.56	.29	.30	--
Global	4.08	.31	.31	.39	--
Air Force	3.84	.48	.14	.23	--

[a]averaged across dimensions within form

[b]average dimension intercorrelations within forms, sources

[c]Supervisor

validity across sources. These show the correlation between two sources averaged over all items on a form.

Several trends are evident from inspection of Tables 1 through 4. First, mean self ratings tend to be slightly higher than mean ratings from peers and supervisors. For example, for Avionic Communications Specialists, mean self ratings ranged from 3.79 to 4.04 across forms, while supervisor ratings ranged from 3.63 to 3.95 and peer ratings ranged from 3.59 to 3.95. It should be noted that this leniency effect for self ratings is only slight and virtually absent for the Information Systems Radio Operator Specialty.

A second trend is that the average dimension intercorrelation within a form is slightly smaller for self ratings than for supervisors or peers. For example, for the Air Traffic Control Operators (see Table 3), the average for self ratings ranged across forms from .32 to .46, but from .41 to .56 for supervisors and from .43 to .51 for peers. Since the average dimension intercorrelation can be interpreted as an index of halo (Saal, Downey, & Lahey, 1980), the present results suggest that incumbents show a greater awareness of their strengths and weaknesses than do supervisors or peers.

Finally, it can be seen that convergent validity coefficients are greater between peers and supervisors than between incumbents and either other source. For Jet Engine Mechanics, the average correlation across dimensions of the

Air Force wide forms was .24 between incumbents and either peers or supervisors, but was .38 between peers and supervisors.

It should be noted that while these analyses are useful for gauging certain "main effects" (e.g., a small difference in means across sources), they do not address the joint or multivariate effects of measurement facets on ratings. Moreover, there is no way of judging the relative contributions to error of each facet. For example, the contribution to measurement error of ratees being evaluated by different combinations of forms and sources is unknown, as is the size of that error component relative to a simple source effect. Nor is it known whether error variance due to these sources can be adequately treated through multiple operationalizations of the measures. Such issues are addressed immediately below.

G Study Results. Summary G study results for analyses of the rating data are presented in Tables 5 and 6. Recall that two G study analyses were performed, one based on four two-item forms and one based on three forms and the number of scales on the dimensional form. Table 5 presents summary results of the four form analysis, and Table 6 presents summary results of the three form analysis. In each of these, variance components for each effect are presented for all four specialties. G study results for each specialty are presented in Appendix A. Results for the four form analysis for each specialty are displayed in

Table 5. Estimated Variance Components for G Study of Rating Variables with Four Forms

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.151	.120	.118	.133
Sources (s)	.015	.015	.036	.001
Forms (f)	.001	-.001	-.017	-.009
Items within f (i:f)	.015	.031	.040	.025
ps	.186	.173	.208	.173
pf	-.003	-.030	-.009	.021
sf	.001	-.008	.000	.003
psf	.016	-.018	.010	.036
p(i:f)	.057	.106	.066	.089
s(i:f)	.004	.019	.000	.002
ps(i:f)	.293	.330	.285	.306

Tables A-1 to A-4, while analyses for the three form design are shown in Tables A-5 to A-8. The within specialty results also display degrees of freedom, mean squares, estimated variance component estimates, and confidence intervals about those estimates for each effect in a design. The confidence intervals indicate the

Table 6. Estimated Variance Components for G Study of Rating Variables with Three Forms

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.100	.095	.084	.106
Sources (s)	.020	.001	.025	-.001
Forms (f)	.033	.034	.016	.049
Items within f (i:f)	.022	.021	.020	.017
ps	.278	.179	.189	.157
pf	.022	.016	.028	.029
sf	.001	.000	-.002	.001
psf	.053	.041	.044	.040
p(i:f)	.057	.069	.048	.031
s(i:f)	.003	.008	.008	.006
ps(i:f)	.290	.291	.339	.303

precision in estimation of the population values of variance components, given the sample size and design complexity. The confidence intervals are based on the ratio of the estimated variance component to its standard error and were calculated from procedures detailed by Satterthwaite (1941, 1946).

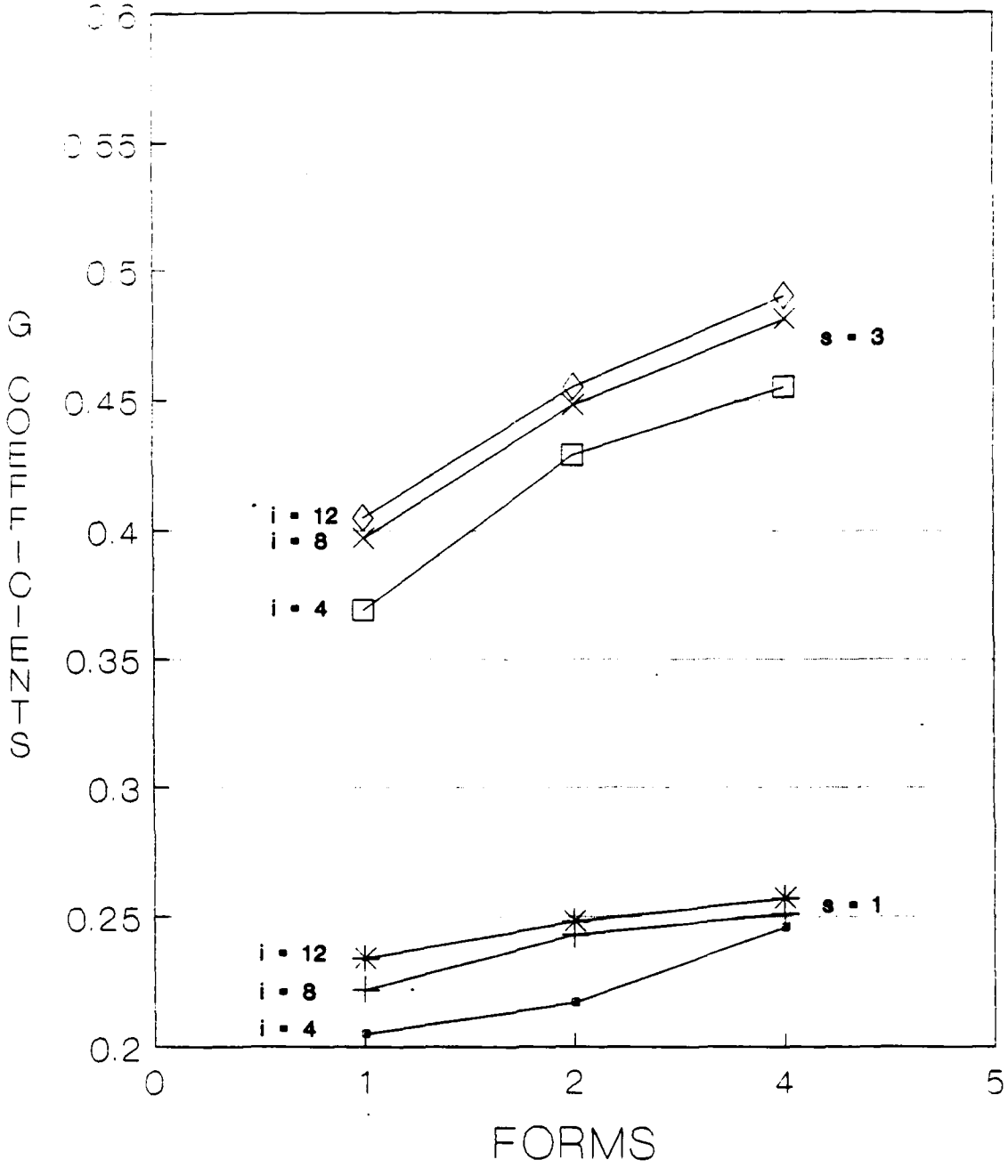
In comparing Tables 5 and 6, it is apparent that the results for the three form and four form analysis are quite

similar. It should also be noted that the results are quite comparable over occupational specialties. For example, looking at the estimates from the four form analyses in Table 5, it can be seen that the largest variance component in each specialty was the residual term ($\underline{\sigma}^2_{ps(i:f)}$) and that the size of this effect ranged from only .285 for Air Traffic Control Operators to .347 for Information Systems Radio Operators. Likewise, the $\underline{\sigma}^2_{ps}$ term is the second largest estimate in each design, ranging from only .139 to .201. The $\underline{\sigma}^2_{p}$ term, universe score variance, is the third largest term for all specialties but Information Systems Radio Operators, and ranges from only .122 to .140. Similar narrow ranges can be seen for the $\underline{\sigma}^2_{pf}$ and $\underline{\sigma}^2_{r(i:f)}$ terms. Only a few terms show considerable variation across specialties. For the four form analyses, the main effect for rater sources, $\underline{\sigma}^2_{s}$ is near zero in three specialties, but substantially larger for Air Traffic Control Operators. Table 3 indicates that this was largely due to low mean supervisory ratings. For both analyses, $\underline{\sigma}^2_{i:f}$ is substantially larger for Information Systems Radio Operators than for other specialties, indicating that items on forms for that specialty were not equivalent in difficulty.

D Study Results. D study analyses of the rating data were based on analyses of the three-form analyses. Results of these analyses are presented by specialty, in Tables A-9 to A-12 of Appendix A and in Figures 5 to 8 below. Figures

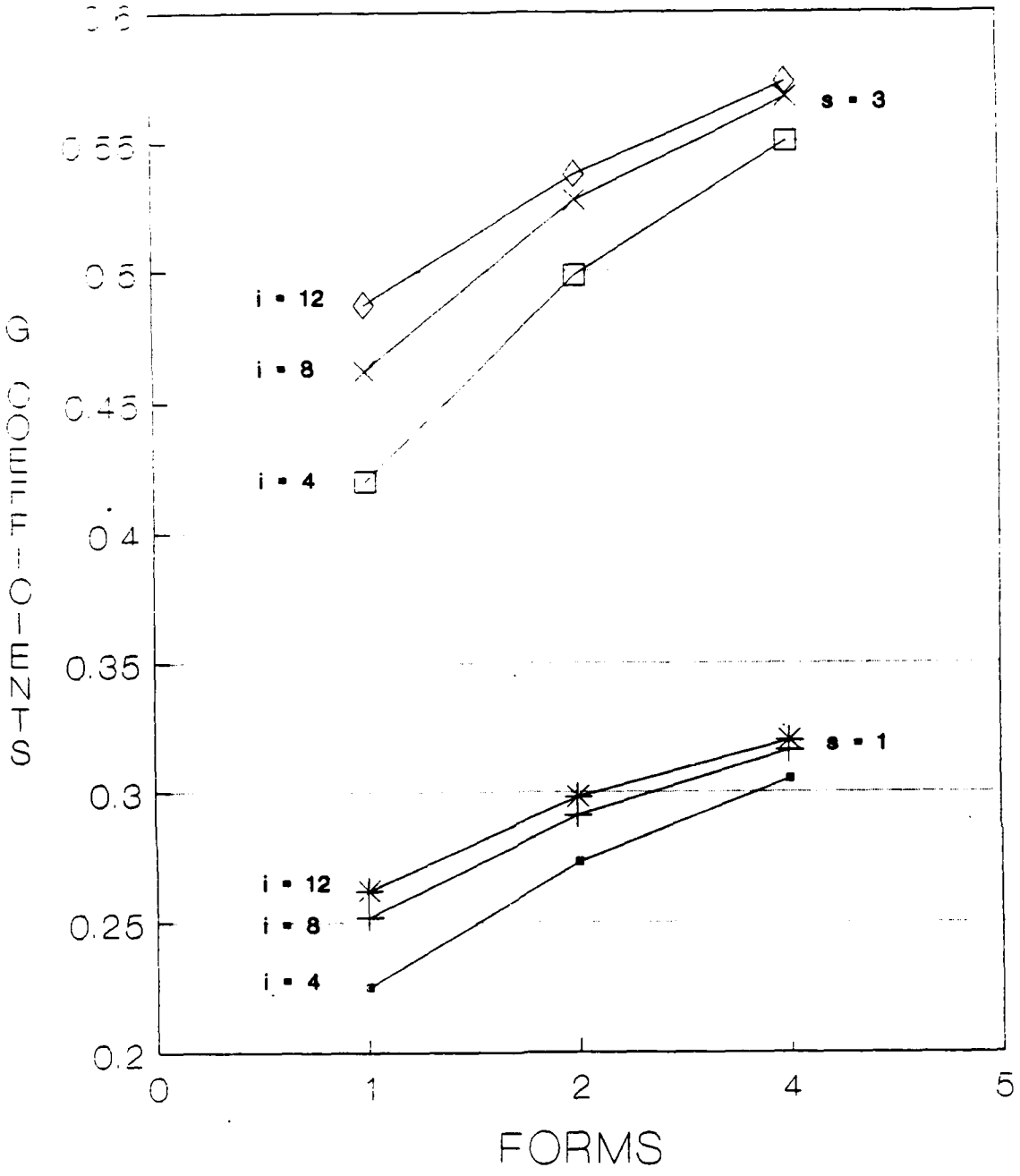
5 through 8 display relationships of generalizability coefficients for relative decisions ($\underline{\epsilon}P^2$) as a function of changes in the number of raters, forms and items within forms. This is the generalizability coefficient ($\underline{\theta}$) for relative decisions. G coefficients for absolute decisions were smaller, but showed similar patterns over changes in D study conditions. The generalizability coefficient represents the proportion of observed score variance which is attributable to universe score variance or individual differences. In each instance, a generalizability coefficient is based on the assumption that scores are those averaged over D study conditions. The curves shown in Figures 5 through 8 summarize results presented in Appendix A Tables A-9 to A-12. By interpolating between points on a curve in any of these figures, decision-makers can estimate the generalizability of performance measures under various conditions not represented by the analyses shown in the tables. Inspection of the pattern of the curves gives additional insight into the relative importance of various sources of variance.

For example, the relatively large effect of the source-by-ratee interaction ($\underline{\sigma}^2_{ps}$) is evident in the figures by the positions of the curves for scores averaged over three sources compared to the curves for any single source. Increasing the number of rater sources is the single best way of increasing the generalizability of the performance ratings. For all four specialties, the



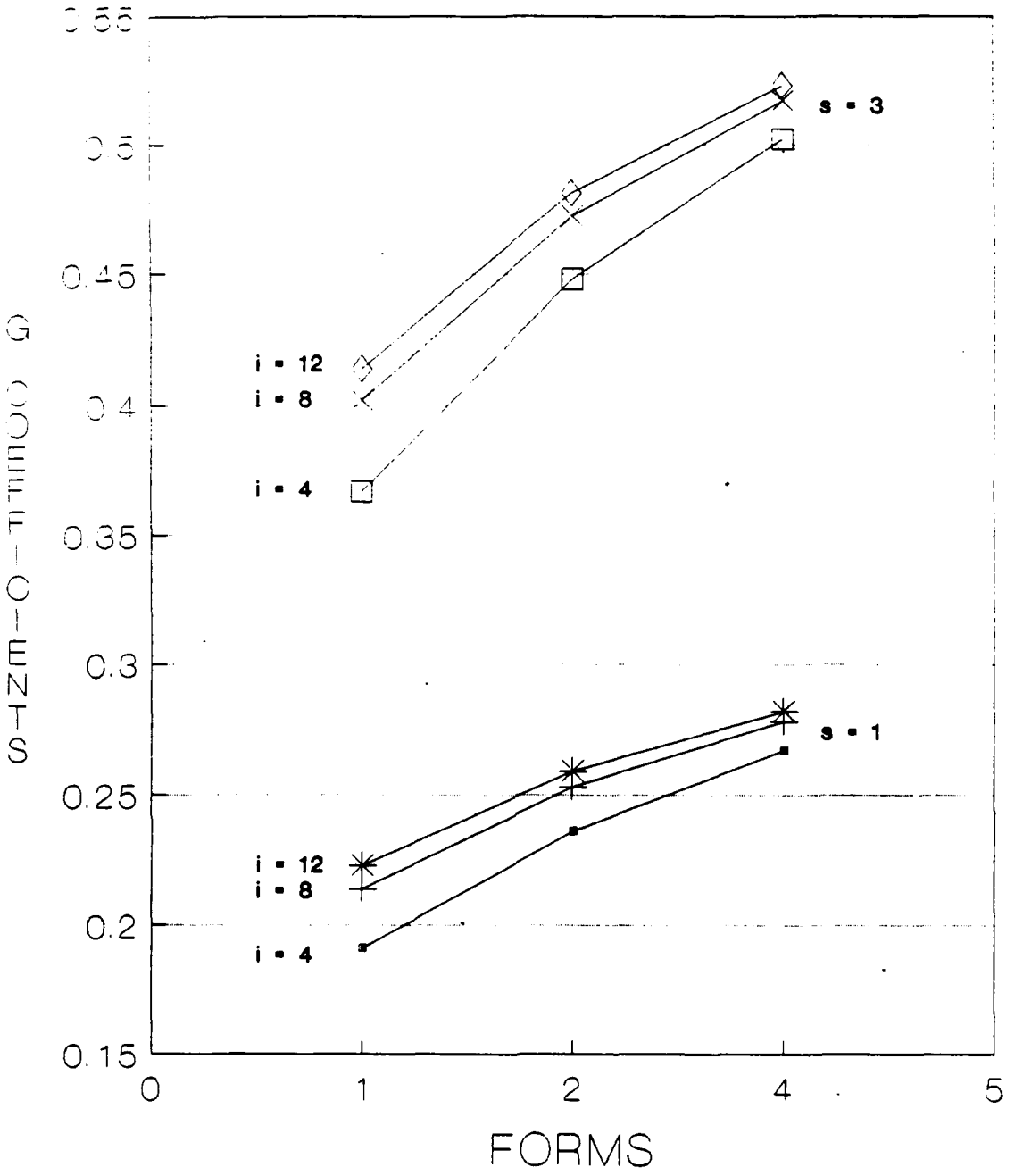
Note. i = Number of Items, s = Number of Sources

Figure 5. Generalizability Coefficient Curves for D Study of Jet Engine Mechanics.



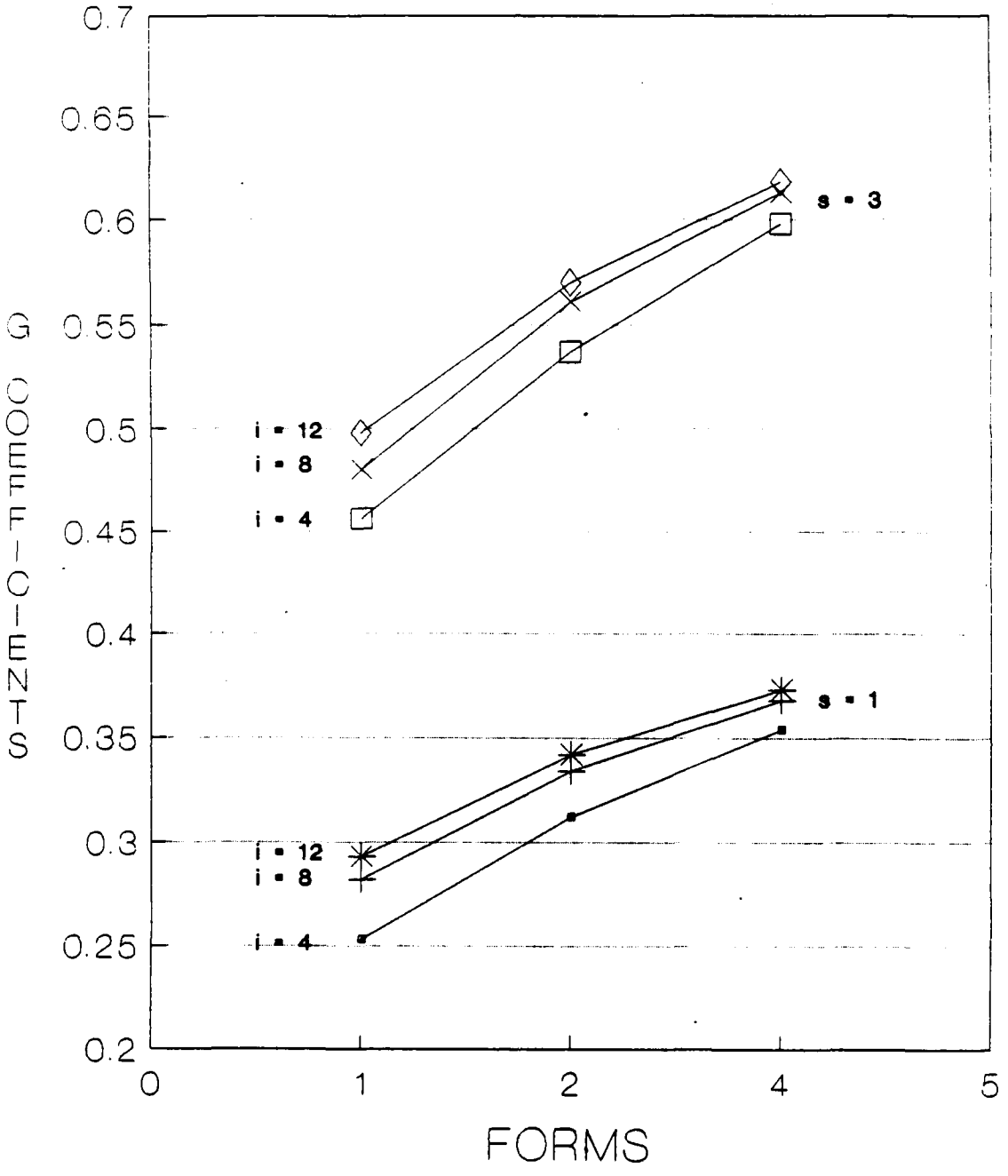
Note. i = Number of Items, s = Number of Sources

Figure 6. Generalizability Coefficient Curves for D Study of Avionic Communications Specialists.



Note. i = Number of Items, s = Number of Sources

Figure 7. Generalizability Coefficient Curves for D Study of Air Traffic Control Operators.



Note. i=Number of Items, s=Number of Sources

Figure 8. Generalizability Coefficient Curves for D Study of Information Systems Radio Operators.

generalizability coefficient for a single 12-item (e.g., task-level) form and three raters is greater than that for a single rater using four eight-item forms.

Looking at the curves for scores averaged over three sources, it can be seen that for all specialties, additional appreciable gains in generalizability can be achieved by using two rating forms instead of one, but somewhat smaller increases in G coefficients result as the number of forms is increased from two to four. For example, for Avionic Communications Specialists, $\underline{\epsilon P^2}$ increases from .252 to .291 as a second 8-item form is added, but only from .291 to .316 as two additional 8-item forms are averaged. Similarly, within levels of rater sources and forms, appreciable gains in generalizability result from increasing the number of items per form from 4 to 8, but not necessarily from 8 to 12. Also, the greater the number of forms, the less the impact of adding items per form. For example, for Air Traffic Control Operators, $\underline{\epsilon P^2}$ increases from .191 to .192 to .223 as the items on a single form are increased from 4 to 8 to 12, but only from .267 to .278 to .281 as items on four forms are added in the same increments. In general, the effects on generalizability coefficients in changes of the number of conditions for each facet are quite consistent across the four specialties. There is some variability in the generalizability coefficients themselves. Regardless of D study conditions, G coefficients for the Jet Engine

Mechanics are smaller than those for the other three specialties. Re-inspection of the variance components in Tables 5 and 6 reveal that these lower generalizability coefficients are primarily because of the relatively high estimate for σ^2_{ps}. Because scores cannot be averaged over more than three sources, the σ^2_{ps} term cannot be sufficiently reduced through addition of measurement conditions to improve generalizability coefficients to satisfactory levels. In contrast, generalizability coefficients for the other three specialties are somewhat larger and quite similar to each other.

Within Source Analyses

Because of the large effect for the interaction of persons and sources, a set of secondary analyses were performed within each rater source for each specialty. In these analyses, facets of interest were forms and items within forms. All analyses employed the three-form design. Both G and D study results for these analyses are displayed in Table 7. A D study coefficient is presented only for a single condition -- ratings on a single 8-item form.

Again, the results were marked by consistency across specialties, for both estimated variance components and generalizability coefficients. The largest source of variance was typically the interaction of persons and items within forms ($\sigma^2_{p(i:f)}$), a term confounded within random

Table 7. G and D Study Results for Within Source Analyses

Source: Effect	JEM σ^2	ACS σ^2	ATC σ^2	ISRO σ^2
Self				
Persons (p)	.192	.161	.218	.219
Forms (f)	.035	.014	.011	.030
Items within forms (i:f)	.025	.034	.021	.019
pf	.053	.030	.038	.073
p(i:f)	.351	.415	.376	.314
ϵP^2 when: f=1, i:f=8	.666	.665	.720	.660
Supervisor				
Persons (p)	.375	.275	.312	.289
Forms (f)	.026	.038	.014	.062
Items within forms (i:f)	.026	.026	.029	.035
pf	.097	.069	.103	.063
p(i:f)	.346	.420	.400	.373
ϵP^2 when: f=1, i:f=8	.728	.694	.671	.726
Peer				
Persons (p)	.265	.357	.291	.282
Forms (f)	.047	.051	.019	.056
Items within forms (i:f)	.024	.017	.031	.015
pf	.077	.020	.075	.072
p(i:f)	.350	.328	.387	.314
ϵP^2 when: f=1, i:f=8	.687	.853	.703	.716

error (σ^2_e). Variance due to individual differences, σ^2_p was also substantial for each source within each specialty, while all other sources of variance were negligible.

In contrast to the prior results, fairly large D study generalizability coefficients were obtained, even under less rigorous measurement specifications (i.e., a single 8-item form). Generalizability coefficients under these conditions ranged from .660 to .853 across sources and specialties. Within the Jet Engine Mechanic and Information Systems Radio Operator specialties, the largest generalizability coefficient was found for the supervisory ratings ($\underline{\epsilon P}^2$ = .728, .726 respectively), while for Avionic Communication Specialists the largest coefficient was found for peer ratings ($\underline{\epsilon P}^2$ = .853), and for Air Traffic controllers the largest coefficient was found for self ratings ($\underline{\epsilon P}^2$ = .720).

G Study Results, WTPT Data

Results of the G study analyses across specialties are presented in Tables 8 (for the crossed design) and 9 (for the nested design). Tables A-13 through A-20 in Appendix A display mean squares, variance components, and confidence intervals for each effect in both designs, shown separately by specialty.

Looking first at results for the crossed design, there is considerably greater variance across specialties than was seen with the WTPT data. For example, variance due to individual differences, σ^2_p, ranged from .006 for Avionic

Table 8. Estimated Variance Components for G Study of WTPT Variables with Tasks, Methods Crossed

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.008	.006	.007	.032
Method (m)	.013	.014	.000	.000
Tasks (t)	.000	.016	.008	.007
Items within t (i:t)	.000	.017	.010	.005
mt	.001	.000	.000	.000
pm	.002	.007	.001	.000
pt	.008	.025	.034	.028
p(i:t)	.009	.032	.073	.012
pmt	.012	.008	.007	.020
m(i:t)	.029	.014	.009	.002
pm(i:t)	.127	.074	.065	.052

Communications Specialists to .032 for Information Systems Radio Operators. Likewise, the residual term was considerably larger in the Jet Engine Mechanic ($\sigma^2_{pm(i:t)}$ = .127) than in the other three specialties. The σ^2_{pm} and σ^2_{pmt} terms were relatively small and consistent across specialties, but considerable variation in estimates was found for the σ^2_{pt} and $\sigma^2_{p(i:t)}$ terms. The estimate for

Table 9. Estimated Variance Components for G Study of WTPT Data with Tasks Nested in Methods

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.008	.013	.007	.029
Methods (m)	.013	.001	-.001	-.001
Tasks within m (t:m)	.003	.014	.012	.008
Items within t w/in m (i:t:m)	.020	.030	.032	.009
pm	.001	-.001	-.002	-.003
p(t:m)	.019	.032	.018	.051
p(i:t:m)	.144	.108	.128	.080

the person by task interaction was near zero for Jet Engine Mechanics, but substantially larger in the other three specialties. This indicates that incumbents in these latter three specialties were differentially ordered on performance, depending on the task. The greatest variability was found for the interactions of persons and items nested with tasks. This term was again near zero for Jet Engine Mechanics, substantially larger for Avionic Communication Specialists and Information Systems Radio Operators, and larger yet for Air Traffic Control Operators. In absolute terms, the estimated variance

component $\underline{\sigma}^2_{p(i:t)}$ for Avionic Communication Specialists and Information Systems Radio Operators was about five times greater and the estimate for Air Traffic Control Operators 15 times greater than the corresponding estimate for Jet Engine Mechanics.

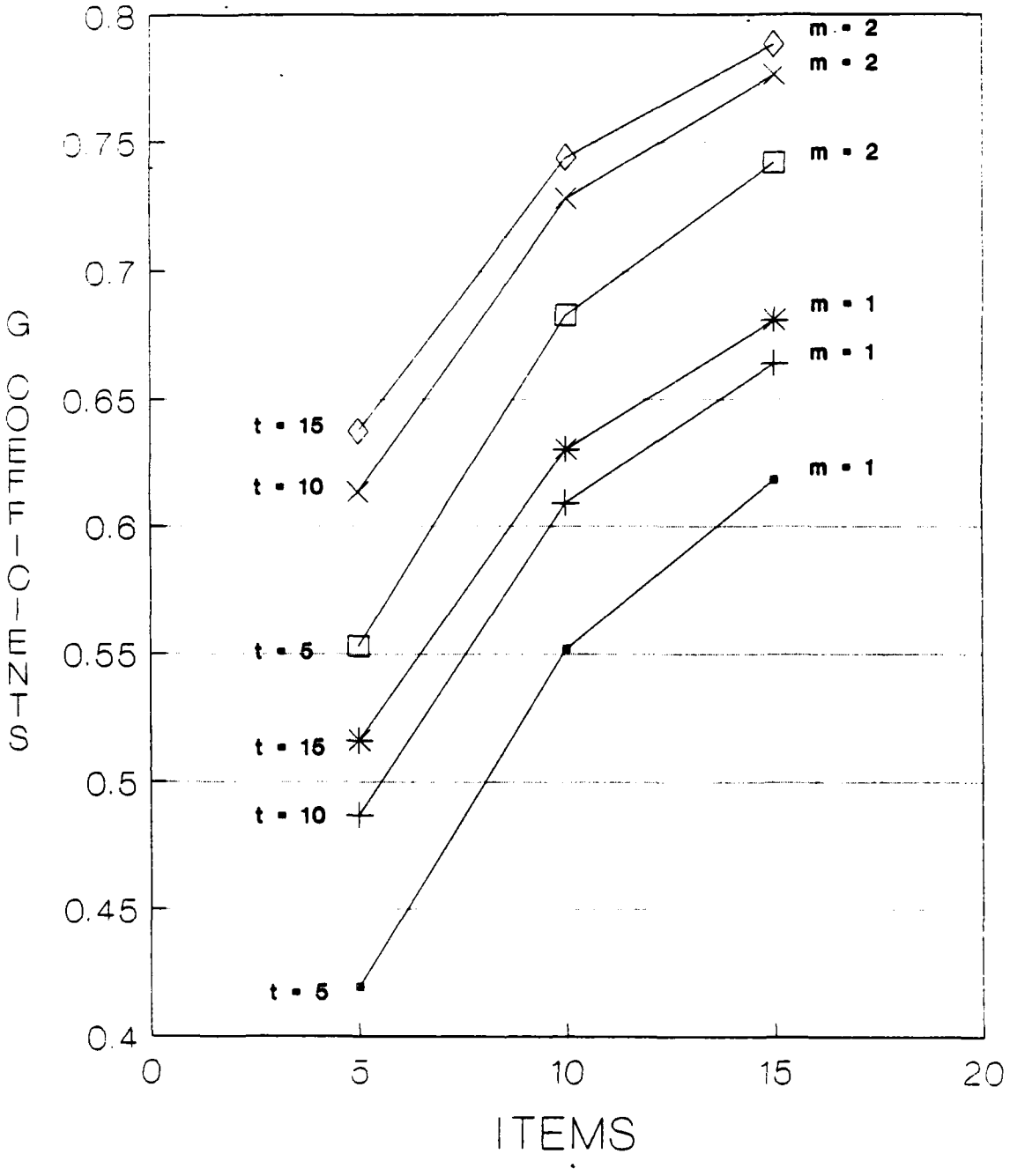
Results for the design with tasks nested in methods were similar to those of the crossed design There was considerable variation across jobs in $\underline{\sigma}^2_{t:m}$ and $\underline{\sigma}^2_{i:t:m}$, but little variation in $\underline{\sigma}^2_{pm}$. These low variance components for the person-by-method interaction indicated that incumbents were not differentially ordered by their performance on the two WTPT methods (hands-on, interview). The residual term, $\underline{\sigma}^2_{p(i:t:m)}$ was the largest variance component for each specialty, though the values of this term varied over specialty. Finally, there was also considerable variation in the $\underline{\sigma}^2_{p(t:m)}$ term, with estimates being substantially lower in the Jet Engine Mechanic and Air Traffic Controller specialties than in the other two AFSs. Thus, only in these two specialties were incumbents not differentially ranked by particular tasks.

D Study Results, WTPT Data

D study analyses were based on the crossed design, since this design permitted assessment of a greater number of effects. D study results for each specialty are displayed graphically in Figures 9 through 12, and in tabular form in Tables A-21 through A-24 in Appendix A.

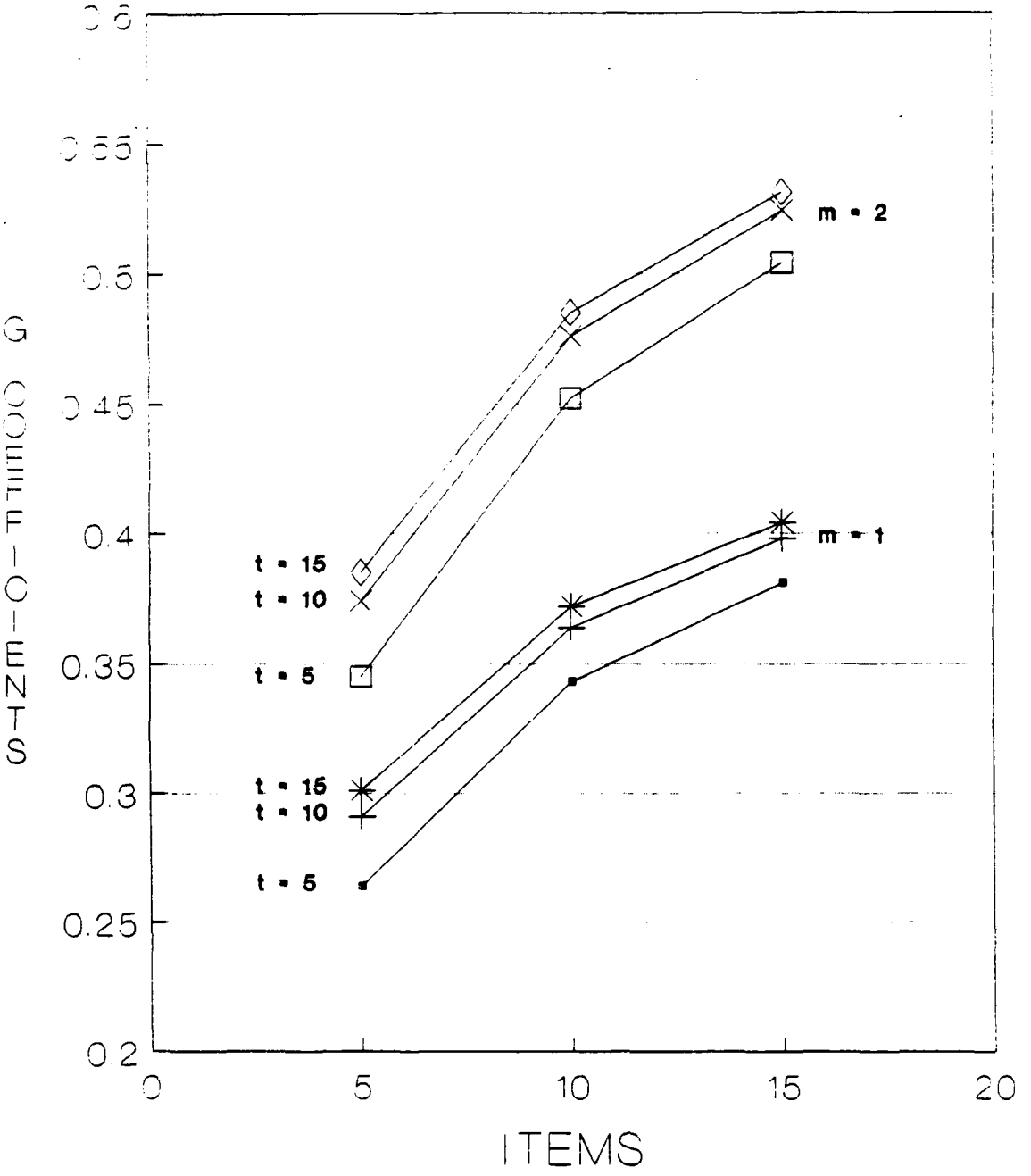
Unlike the D study results for the rating data, changes in specifications of measurement conditions produced considerable variations in the resulting generalizability curves. Just as increasing the number of rating sources had the largest effect on the rating data, using both WTPT methods (i.e., hands-on and interview) had the greatest effect on WTPT generalizability coefficients. In general, scores averaged over both methods using a small number of items and a small number of tasks were more generalizable than scores on a single method with a substantially greater number of tasks or items.

Inspection of Figures 9 through 12 reveals that the greatest levels of generalizability were obtained for Information Systems Radio Operators and Jet Engine Mechanics. For the latter, generalizability coefficients above .750 would only be obtained under fairly extensive measurement conditions - 15 tasks, each with 10 steps, assessed by both hands-on and interview formats. In contrast, generalizability coefficients above .800 for Information Systems Radio Operators result from a variety of conditions. Under the most extensive measurement conditions studied (two methods, 15 tasks, 15 items), $\underline{\epsilon}P^2$ equaled .922. When at least 10 tasks are used, generalizability coefficients over .750 resulted with either one or two methods and 5, 10, or 15 items. Only averaging scores over five tasks produces generalizability coefficients less than .750.



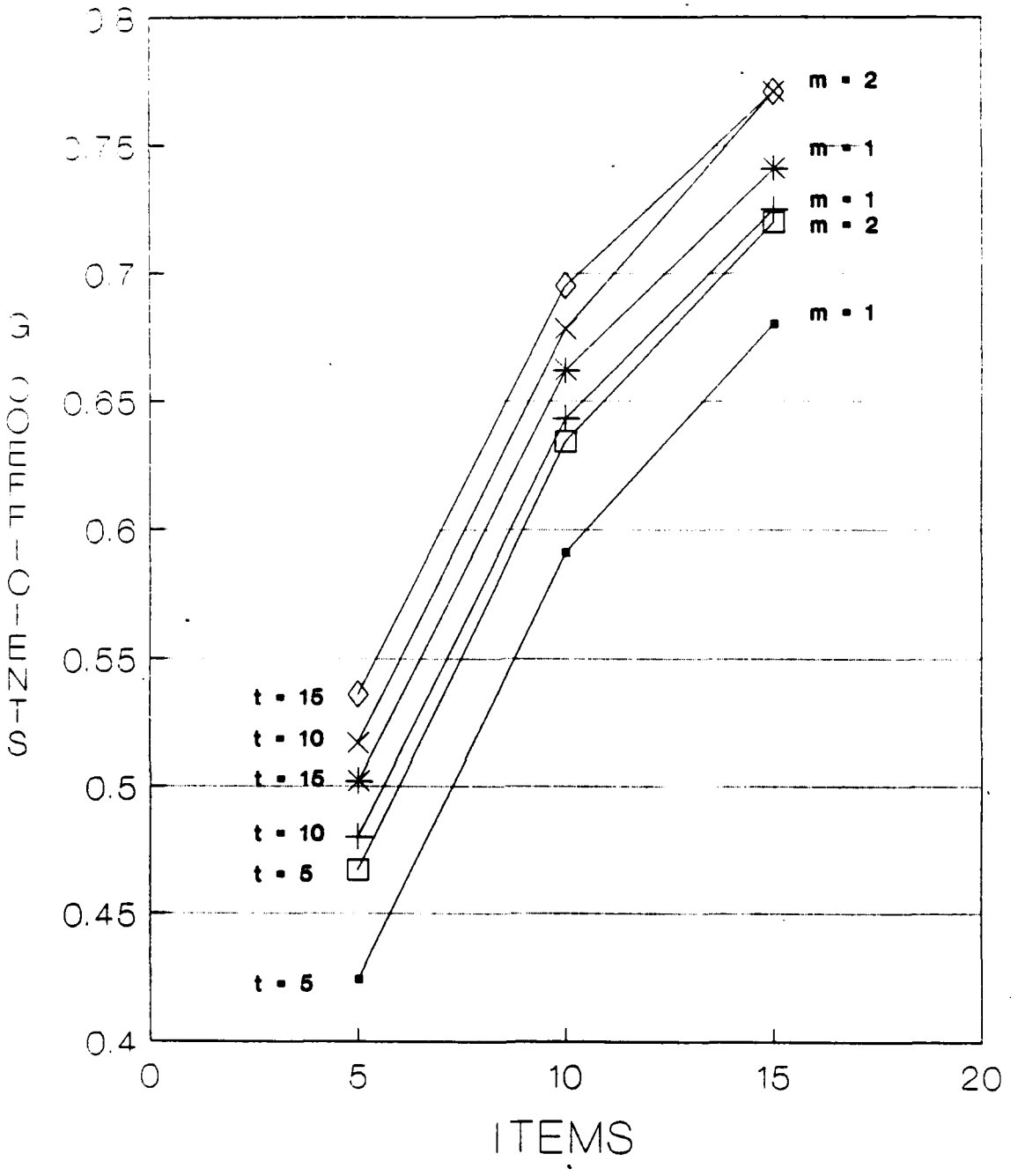
Note. t = Number of Tasks, m = Number of Methods

Figure 9. Generalizability Coefficient Curves for WTPT D Study of Jet Engine Mechanics.



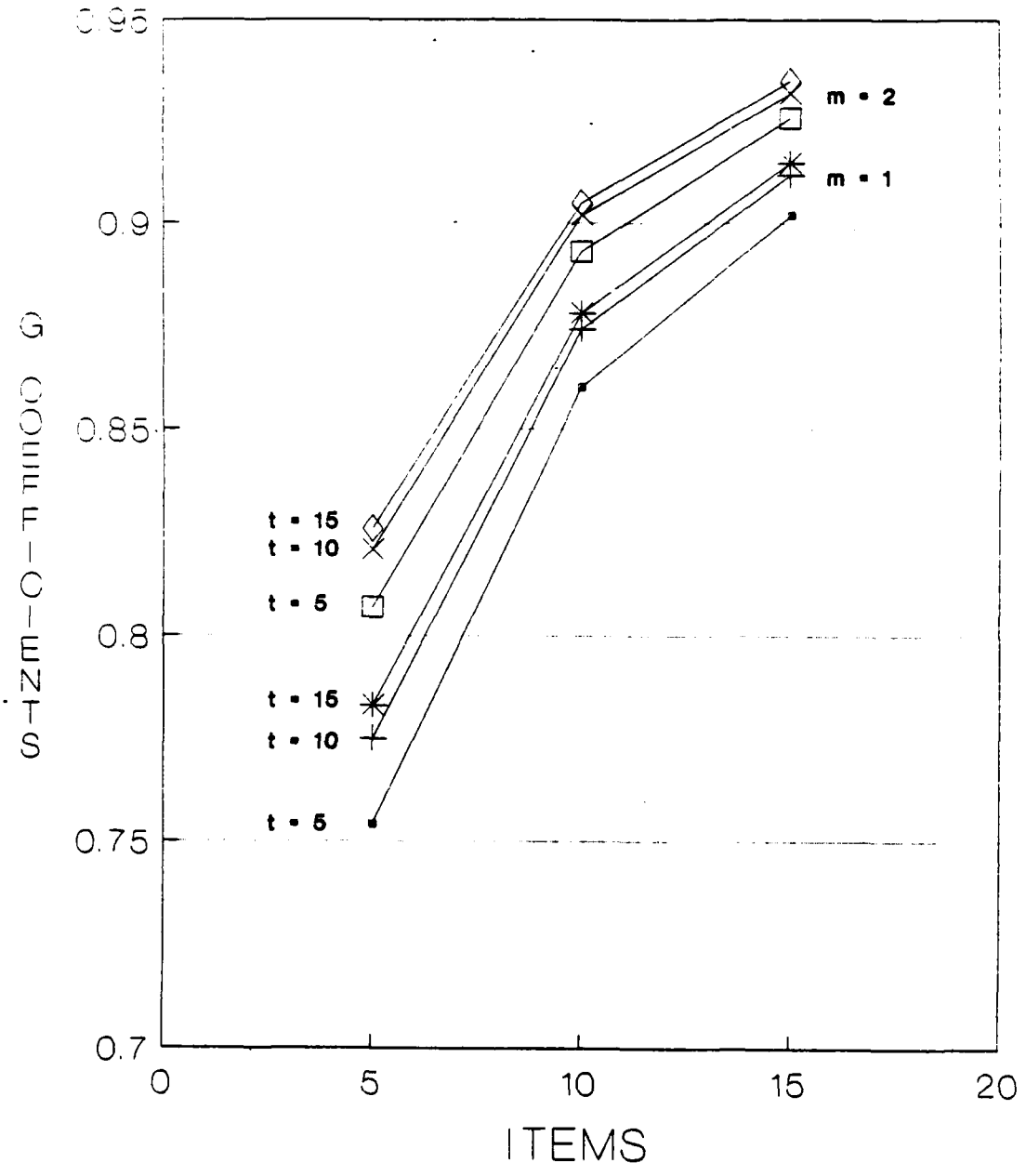
Note. t = Number of Tasks, m = Number of Methods

Figure 10. Generalizability Coefficient Curves for WTPT D Study of Avionic Communications Specialists.



Note. t = Number of Tasks, m = Number of Methods

Figure 11. Generalizability Coefficient Curves for WTPT D Study of Air Traffic Control Operators.



Note. t = Number of Tasks, m = Number of Methods

Figure 12. Generalizability Coefficient Curves for WTPT D Study of Information Systems Radio Operators.

Generalizability coefficients were considerably lower in the other two specialties. The lowest levels of generalizability occurred for Avionic Communications Specialists. Even with scores averaged over two methods,

15 tasks, and 15 items, $\underline{\epsilon P}^2$ equaled only .531. Generalizability levels were somewhat higher for the Air Traffic Control Operators, with $\underline{\epsilon P}^2$ equal to .698 under the most stringent conditions specified. It is clear that for this specialty, the WTPT should be constructed with as many items and tasks as feasible.

G and D Study Results, Substitutability Design

G study estimated variance components, as well as D study estimates of $\underline{\epsilon P}^2$ of the substitutability design are presented in Tables 10 through 13. The substitutability design reflects variability in individuals' performance scores across proficiency ratings, hands-on tests, and interview tests. D study estimates are presented for two sets of measurement conditions: A single method of assessing 15 tasks and scores averaged over all three methods, each assessing 15 tasks. In Table 10, proficiency ratings are first averaged over all three sources before being compared to the WTPT scores. In Tables 11 to 13, ratings are analyzed separately by source (self, supervisor, and peer, respectively).

In no instance are performance scores generalizable over the three evaluation methods. The highest levels of generalizability occur for the design which includes ratings averaged over sources. Even here, when only a single method is used, $\underline{\epsilon P}^2$ ranged from only .112 to .369. Thus, at best, only a little over a third of the observed

Table 10. G and D Study results for Substitutability Design with Ratings Averaged Across Sources

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.010	.012	.010	.030
Methods (m)	3.446	3.302	3.201	4.151
Tasks (t)	-.003	.005	.007	.002
mt	.039	.017	.051	.020
pm	.054	.053	.070	.046
pt	.000	.000	.006	.002
pmt	.091	.115	.097	.081
ϵP^2 when:				
m = 1, t = 15	.140	.166	.112	.369
m = 3, t = 15	.327	.373	.274	.637

variance in individuals' scores can be attributed to universe score variance (or individual differences). In this design, when scores are further averaged over the three assessment methods (and 15 tasks), generalizability coefficients still range from only .274 to .637. While the latter coefficient (for Information Systems Radio Operators) approaches a respectable level, it should be noted that coefficients for the other three specialties are

Table 11. G and D Study Results for Substitutability Design with Self Ratings

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.005	.013	.011	.035
Methods (m)	3.691	3.436	3.505	4.097
Tasks (t)	-.003	.005	.007	.001
mt	.049	.019	.066	.022
pm	.088	.075	.095	.085
pt	.001	.000	.007	.000
pmt	.176	.207	.196	.156
ϵP^2 when:				
$m = 1, t = 15$	.043	.129	.088	.266
$m = 3, t = 15$	.120	.307	.222	.520

all less than .400. When ratings are analyzed separately by source, generalizability coefficients under either set of D study conditions are even smaller. Interestingly, there is no tendency for generalizability coefficients to be consistently higher or lower for any single source. Looking at the individual variance components, it is clear that the low generalizability coefficients are the result of large values for the σ^2_{pm} and σ^2_{pmt} terms. The

Table 12. G and D Study Results for Substitutability Design with Supervisor Ratings

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.016	.012	.007	.031
Methods (m)	3.202	3.196	2.801	4.219
Tasks (t)	-.002	.002	.006	.002
mt	.033	.021	.042	.023
pm	.130	.126	.149	.086
pt	-.002	.002	.006	.002
pmt	.144	.188	.181	.137
ϵP^2 when:				
$n_m = 1$, $n_t = 15$	.104	.076	.044	.244
$n_m = 3$, $n_t = 15$	.259	.198	.120	.491

σ^2_{pt} term are near zero in all analyses, indicating that persons were similarly ranked on all tasks. However, σ^2_{pm} and σ^2_{pmt} are considerably larger, regardless of analysis. Though at the G study level, σ^2_{pmt} is larger than σ^2_{pm}, σ^2_{pM} is typically larger at the D study level, since σ^2_{pMT} can be reduced by averaging scores over methods and tasks. Since it is the D study estimates of σ^2_{p}, σ^2_{pM}, σ^2_{pT}, and σ^2_{pMT} which are used to compute ϵP^2, it is clear that it is the differential ordering of persons over methods which produces the low generalizability coefficients.

Table 13. G and D Study Results for Substitutability Design with Peer Ratings

	Job:			
	JEM	ACS	ATC	ISRO
Effect	σ^2	σ^2	σ^2	σ^2
Persons (p)	.010	.009	.009	.025
Methods (m)	3.338	3.196	3.310	4.139
Tasks (t)	-.003	.008	.006	.002
mt	.046	.014	.049	.019
pm	.107	.096	.148	.097
pt	.000	.004	.004	.002
pmt	.144	.175	.170	.153
ϵP^2 when:				
$n_m = 1$, $n_t = 15$	.077	.079	.056	.189
$n_m = 3$, $n_t = 15$	.200	.205	.150	.411

The high estimates for σ^2_m indicate large mean differences between methods and are an artifact produced by a 5-point scale used for the ratings and a 1-point scale used for the two WTPT methods. (A 1-point scale resulted from scores on tasks computed as the average of a number of dichotomous -- correct/incorrect -- items).

IV. DISCUSSION

The purpose of the present investigation was to apply generalizability theory to the data collected on the Air

Force Performance Measurement Project in order to address the following issues: The psychometric adequacy of the proficiency ratings, the psychometric adequacy of the Walk-Through performance tests, and the extent to which the ratings are acceptable surrogates for the WTPT. Also of interest are whether results are consistent across specialties, and whether particular measurement technologies can be reduced in scope without comprising the dependability of scores. Each of the issues is addressed below, along with recommendations regarding the performance measurement project.

Psychometric Quality of Performance Ratings

Evidence for the psychometric quality of the performance ratings comes from G and D study results within each occupational specialty. Cardinet, Tourneur, and Allal (1976) recommended .80 as a minimally acceptable level for generalizability coefficients. Given this value, the generalizability levels of proficiency ratings for relative decisions are inadequate in each specialty, regardless of the measurement conditions specified. However, it can be argued that the recommendations of Cardinet et al. were made principally for paper-and-pencil tests, and it is logical to expect G coefficients for rating systems to be lower. At the least, the suitability of any generalizability coefficient should be interpreted within the context of results from similar studies.

Given these qualifications, it is reasonable to be somewhat optimistic about the fidelity of the proficiency ratings. For three of the specialties, generalizability coefficients are about .70 when scores are averaged over three sources, at least two forms, and at least 8 to 10 items. Generalizability coefficients for the fourth specialty (Air Traffic Control Operators) are only slightly lower. This indicates that under such measurement conditions, about 70% of the observed variance in scores can be attributed to individual differences. These generalizability coefficients are about the same as, or greater than, coefficients reported in similar rating studies by Littlefield, Murrey, and Garman (1977), McHenry et al. (1987), and Webb and Shavelson (1987). Further, they are higher than typical inter-rater reliability estimates (King, Hunter, & Schmidt, 1980).

Brennan and Kane (1979) noted that inspection of individual variance components is often more illuminating than summary coefficients. It is clear from Tables 5 and 6 that the largest sources of error variance are the ratee-by-source and ratee-by-source-by-items within forms interactions ($\underline{\sigma}^2_{ps}$, $\underline{\sigma}^2_{ps(i:f)}$). The latter term also contains undifferentiated error, $\underline{\sigma}^2_{e}$. While this term is largest at the G study level, it can be substantially reduced at the D study level by replication in sources, items, and forms. In contrast, the $\underline{\sigma}^2_{ps}$ term can only be reduced at the D study level by averaging over sources.

Thus, it exerts the largest influence on the G coefficient computations. Since this term is large when ratees are differentially ranked by sources the present results are consistent with hypotheses that sources differ in their opportunities to observe or in their interpretation of behavior (Borman, 1974; Guion, 1966; Klimoski, & London, 1974). Operationally, the implication of these results are that the Air Force should continue collecting and averaging scores over sources to reduce error variance at the D study level and should also consider changes in rater training or rater instructions to increase uniformity across sources and reduce error variance at the G study level. Given that all three sources are used, there seems to be little gain in using more than one or two forms or more than 8 to 10 items per form.

Finally, it is noted that results are very consistent across the four specialties studied. There appears to be little or no variability across jobs in the psychometric characteristics of the rating system. Thus, there is less of a need to continue collecting and assessing rating data in additional specialties, unless attempts are made (and tested) to increase convergence across sources.

Psychometric Quality of WTPT Scores

Evidence of the psychometric quality of the Walk-Through Performance Testing method comes from G and D study results within each occupational specialty. In contrast to the rating data, there is considerably greater variability

across specialties for the WTPT data. Looking first at the D study generalizability coefficients (see Figures 9 through 12), it can be seen that acceptable levels of generalizability are reached under a variety of measurement conditions for the jobs of Jet Engine Mechanic and Information Systems Radio Operator. For the latter job, generalizability coefficients greater than .80 can be achieved even when a single method is used with as few as 10 tasks. When scores are averaged over at least 15 tasks assessed by both the hands-on and interview components, $\underline{\epsilon}P^2$ is about .90. For Jet Engine Mechanics, generalizability coefficients are smaller, and scores must be averaged over at least 15 tasks assessed by both methods to produce generalizability coefficients of .80.

Generalizability coefficients under all conditions are smaller yet for the other two specialties. For Air Traffic Control Operators, G coefficients approach .70 when 15 tasks, 10 items, and two methods are used, but are considerably smaller with fewer tasks or items or a single method. For Avionic Communications Specialists, D study generalizability coefficients are well below .50 under all measurement conditions studied.

Inspection of the G study estimates of variance components reveals that the interaction of persons and tasks ($\underline{\sigma}^2 pt$ or $\underline{\sigma}^2 p(t:m)$) the interaction of persons and items ($\underline{\sigma}^2 p(i:t)$ or $\underline{\sigma}^2 p(i:t:m)$) and the residual term ($\underline{\sigma}^2 pm(i:t)$ or $\underline{\sigma}^2 p(i:t:m)$ all contributed substantial error

variance in different combinations of jobs or designs. Whether tasks were treated as crossed with methods or nested within methods, the interaction of persons and tasks contributed a substantial portion of variance in all specialties but Jet Engine Mechanic. This indicates that persons were differentially ranked in terms of performance on tasks. This could indicate incomplete or inconsistent training in these specialties either within or across bases. For example, if Airman A is adequately trained to perform task 1 but not task 2, while Airman B is trained to perform task 2 but not one, these two Airman will be differentially ranked on these two tasks, even if there are no true differences in job proficiency. It would be interesting to compare the thoroughness and consistency of training of Jet Engine Mechanics to that of the other specialties. An alternative plausible explanation is differential mission requirements which affect individual opportunities to perform. Opportunities to perform tasks may vary by person, within and across bases, and will be reflected in a high person-by-task interaction. This hypothesis could be tested using the frequency of task performance information that has been collected as part of the WTPT data collection.

Sampling of items was also an issue, particularly for Air Traffic Control Operators. The variance component $\sigma^2_{p(i:t)}$ for Air Traffic Control Operators was over twice as large as that for Information Systems Radio Operators or

Avionic Communication Specialists, and over 10 times as large as that for Jet Engine Mechanics. This variance component reflects differential ordering of persons on items or steps within individual tasks. One reason this variance component may be so large for Air Traffic Control Operators is that individual items constituting tasks in fact represent different underlying dimensions of performance. For example, the WTPT tasks require incumbents to demonstrate proficiency at both the skillful reaction of complex landing operations and the completion of forms or following of precise procedures. It is possible that individuals skillful at the former task are less able to demonstrate proficiency at the latter tasks. Thus, while their task summary scores would be similar to persons who are better at following the book but less capable at the technical aspects of the job, they would be differentially ranked at the item level.

The residual terms are somewhat high, especially for Jet Engine Mechanics and for Air Traffic Control Operators when tasks are treated as nested within methods. For the latter job, this residual term includes the confounding of the $\underline{\sigma}^2_{p(i:t:m)}$ and $\underline{\sigma}^2_e$ terms. Since the interaction of persons and items was large for this specialty in the crossed design, it is safe to assume that the $\underline{\sigma}^2_{p(i:t:m)}$ term accounts for much of the variance in the residual term for the reasons speculated above. For Jet Engine Mechanics though, the residual term is large in both designs. For

the crossed design, the residual term confounds $\sigma^2_{pm(i:t)}$ and σ^2_e. Since other terms containing the interaction of persons and methods in this design are very small (σ^2_{pm} and σ^2_{pmt}), it can be reasonably assumed that it is the effects of σ^2_e which results in the extremely high residual term for this specialty. Undifferentiated error includes both random error and other systematic effects not included in the design. For example, if persons were differentially ranked by test administrators, or persons from various bases were differentially ranked, these effects would be reflected by the residual term, but could not be assessed by the present design. At best, Air Force decision makers could intuitively judge whether it is plausible to assume that administrators, bases, or other systematic effects were more problematic with the Jet Engine Mechanic specialty than others. On the other hand, since this was the first specialty in which the WTPT was designed and administered, decision makers may also wish to judge whether it is likely that there was greater random error introduced by the newness of procedures.

It should also be noted that the variance component for the persons-by-methods interaction (σ^2_{pm}) was extremely small in both designs, for all four specialties. This means that test-takers were ranked the same whether they were actually performing the task or merely describing it. Thus, the interview format is a more than acceptable substitute for the more time-consuming hands-on component.

Finally, the variability in variance components and generalizability coefficients across specialties is re-emphasized. It would be unwise to discontinue attempts to design and implement Walk-Through Performance Testing in other specialties until clearer patterns of results are uncovered.

Ratings as Surrogates for WTPT Scores

Evidence of the adequacy of proficiency ratings as surrogates of the WTPTs comes from G and D studies of the substitutability design. Regardless of whether scores are averaged across sources, or considered for each source by itself, there is very little convergence between ratings and WTPT scores. Even averaging scores over many tasks does not improve the *generalizability* of scores over the two evaluation methods. Thus, task proficiency ratings are not adequate surrogates for the WTPT.

One question which follows is which set of scores is the more trustworthy. Under normal measurement conditions, the generalizability analyses discussed above indicate that the performance ratings are more dependable for Avionic Communications Specialists and Air Traffic Control Operators, but that WTPT scores are more dependable for Jet Engine Mechanics and Information Systems Radio Operators. Such conclusions are tempered by the confidence one has that all measurement conditions which might affect scores were included in analyses of the ratings and WTPT scores. For example, if test administrators did contribute

significant error variance to WTPT scores, designs which permitted estimation of such effects could have resulted in superior generalizability coefficients for WTPT tests in all specialties. At present though, there appears to be no reason to favor one methodology over the other and the wisest course of action would seem to be to continue using both sets of scores in decisions.

Recommendations

These recommendations for the Air Force JPM project are made under two sets of assumptions. First, it is assumed that there is a desire to continue collecting performance data in additional specialties for purposes of understanding the psychometric/measurement properties of the various evaluation methods. Secondly, it is assumed that there is an intention to continue collecting performance data in additional specialties for purposes of validating the Armed Services Vocational Aptitude Battery (ASVAB) and evaluating technical training school performance.

1. There appears to be little utility in collecting additional information on proficiency ratings for purposes of understanding their psychometric quality. Results to date are remarkably consistent across the specialties already studied. The best reason to continue studying ratings data would be to test differences in aspects of scale development or data collection (e.g, variations in rater training programs). From a research perspective, it

would be valuable to continue exploring the differential meaning and validity of ratings by different sources.

2. Proficiency ratings appear to be adequate criteria for validation purposes and the methodology developed in these four specialties should be applied in others as well. It is possible to reduce the number of forms to one or two and maintain current fidelity levels, but ratings should be collected from (and averaged over) all three sources.

3. The WTPTs should be applied in other specialties for both pure research and validation purposes. Additional research is needed because the expected generalizability coefficients or relative size of individual variance components cannot be extrapolated from the data collected to date. As data is collected from additional specialties, it may be possible to draw generalizations on the adequacy of WTPT scores depending on the nature of the job (e.g., mechanical vs. administrative). The WTPT should be used for validation purposes in additional specialties because of evidence that it is more dependable than proficiency ratings in two specialties examined thus far. The extremely high generalizability coefficients in one specialty justifies the efforts the Air Force has applied to this method, though the extremely low coefficients for Air Traffic Control Operators call for caution and much additional research and/or improvements in developmental technologies.

4. It is unwise to consider proficiency ratings as surrogates for the WTPT. Instead, they appear to represent vastly different aspects of the total criterion space. While each methodology is reliable and dependable in and of itself, there is little overlap in the substantive universes assessed by each. Thus, both rating data and WTPT scores should be considered "correct," even though they are essentially unrelated. Other research strategies which emphasize comparing both sets of scores to other indicators or predictors of performance appear to be necessary to understand the latent constructs measured by each (Borman, 1987).

REFERENCES

Borman, W.C. (1974). The ratings of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 12, 105-124.

Borman, W.C. (1987, April). Comments as panelist in M.M. Kavanagh (Chair), A roundtable discussion of research issues in criterion measurement. Annual meeting of the Society for Industrial/Organizational Psychology, Atlanta, GA.

Brennan, R.L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Brennan, R.L., & Kane, M.T. (1979). Generalizability theory: A review. In L.J. Fryans, Jr. (Ed.), Generalizability theory: Inferences and practical applications. San Francisco: Jossey-Bass.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 13, 119-135.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York: Wiley.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Gillmore, G.M. (1979, March). An introduction to generalizability theory as a contributor to evaluation research. Washington University, Seattle: Educational Assessment Center.

Gillmore, G.M. (1983). Generalizability theory: Application to program evaluation. In L.J. Fryans, Jr. (Ed.), Generalizability theory: Inferences and practical applications. San Francisco: Jossey-Bass.

Guion, R.M. (1966). Personnel testing. New York: McGraw-Hill.

Hedge, J.W., & Teachout, M.S. (1986, November). Job performance measurement: A systematic program of research and development (AFHRL-TP-86-37, AD-A174 175). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

King, L.M., Hunter, J.E., & Schmidt, F.L. (1980). Halo in a multidimensional forced choice performance evaluation scale. Journal of Applied Psychology, 65, 507-516.

Klimoski, R.J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59, 445-451.

Kraiger, K. (1989, April). Generalizability theory: An assessment of its relevance to the Air Force job performance measurement project (AFHRL-TP-87-70, AD-A207-107). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Kraiger, K., & Teachout, M.S. (1987, April). Generalizability theory as evidence of the construct validity of ratings. In G. Laabs (Chair), Applications of generalizability theory to military performance measurement. Symposium conducted at the annual meeting of the American Educational Research Association, Washington, DC.

Kraiger, K., & Teachout, M.S. (1990). Generalizability theory as constructed-related evidence of the validity of job performance ratings. Human Performance, 3, 19-35.

Littlefield, J.H., Murrey, A.J., & Garman, R.E. (1977, April). Assessing the generalizability of clinical rating scales. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

McHenry, J.J., Hoffman, R.G., & White, L.A. (1987, April). A generalizability analysis of peer and supervisory ratings. In G. Laabs (Chair), Applications of generalizability theory to military performance measurement. Symposium conducted at the annual meeting of the American Educational Research Association, Washington, DC.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Satterthwaite, F.E. (1941). Synthesis of variance. Psychometrika, 6, 309-316.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2, 110-114.

Searle, S.R. (1971). Linear models. New York: Wiley.

Shavelson, R.J. (July, 1986). Generalizability of military performance measurements: I. Individual performance. Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council, National Academy of Sciences, Washington, DC.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-161.

Webb, N., & Shavelson, R. (1937, April). Generalizability theory and job performance measurement. In G. Laabs (Chair), Applications of generalizability theory to military performance measurement. Symposium conducted at the annual meeting of the American Educational Research Association, Washington, DC.

APPENDIX A:

ADDITIONAL G AND D STUDY RESULTS WITHIN OCCUPATIONAL SPECIALTIES

Table A-1. Estimated Variance Components for Jet Engine Mechanics, Four-form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	206	5.591	.151	$.119<\sigma^2<.199$
Sources (s)	2	27.841	.015	$.008<\sigma^2<.044$
Forms (f)	3	12.147	.001	$.001<\sigma^2<.003$
Items within f (i:f)	4	10.639	.015	$.008<\sigma^2<.044$
ps	412	1.812	.186	$.163<\sigma^2<.214$
pf	618	.477	-.003	$.000<\sigma^2<.000$
sf	6	1.434	.001	$.000<\sigma^2<.002$
psf	1,236	.326	.016	$.009<\sigma^2<.048$
p(i:f)	824	.464	.057	$.046<\sigma^2<.074$
s(i:f)	8	1.086	.004	$.002<\sigma^2<.011$
ps(i:f)	1,648	.293	.293	$.276<\sigma^2<.311$

Table A-2. Estimated Variance Components for Avionic Communication Specialists, Four-Form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	82	4.698	.120	$.082<\sigma^2<.195$
Sources (s)	2	11.809	.015	$.008<\sigma^2<.044$
Forms (f)	3	7.770	-.001	$.000<\sigma^2<.000$
Items within f (i:f)	4	9.987	.031	$.016<\sigma^2<.092$
ps	164	1.678	.173	$.140<\sigma^2<.219$
pf	246	.431	-.030	$.000<\sigma^2<.000$
sf	6	.546	-.008	$.000<\sigma^2<.000$
psf	492	.295	-.018	$.000<\sigma^2<.000$
p(i:f)	328	.648	.106	$.082<\sigma^2<.143$
s(i:f)	8	1.876	.019	$.010<\sigma^2<.054$
ps(i:f)	656	.330	.330	$.302<\sigma^2<.363$

Table A-3. Estimated Variance Components for Air Traffic Control Operators, Four-form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	187	4.944	.118	$.089<\sigma^2<.165$
Sources (s)	2	55.523	.036	$.018<\sigma^2<.104$
Forms (f)	3	3.289	-.017	$.000<\sigma^2<.000$
Items within f (i:f)	4	22.994	.040	$.021<\sigma^2<.116$
ps	374	1.970	.208	$.181<\sigma^2<.243$
pf	561	.451	-.009	$.000<\sigma^2<.000$
sf	6	.191	.000	$.000<\sigma^2<.000$
psf	1,122	.305	.010	$.005<\sigma^2<.030$
p(i:f)	748	.483	.066	$.054<\sigma^2<.084$
s(i:f)	8	.377	.000	$.000<\sigma^2<.001$
ps(i:f)	1,496	.285	.285	$.268<\sigma^2<.302$

Table A-4. Estimated Variance Components for Information System Radio Operators, Four-form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	151	5.295	.133	$.099<\sigma^2<.190$
Sources (s)	2	3.889	.001	$.000<\sigma^2<.003$
Forms (f)	3	2.637	-.009	$.000<\sigma^2<.000$
Items within f (i:f)	4	9.735	.020	$.001<\sigma^2<.006$
ps	302	1.760	.173	$.146<\sigma^2<.207$
pf	453	.725	.013	$.007<\sigma^2<.038$
sf	6	1.411	.003	$.001<\sigma^2<.008$
psf	906	.379	.036	$.024<\sigma^2<.055$
p(i:f)	604	.575	.090	$.073<\sigma^2<.113$
s(i:f)	8	.562	.002	$.001<\sigma^2<.005$
ps(i:f)	1,208	.306	.306	$.295<\sigma^2<.330$

Table A-5. Estimated Variance Components for Jet Engine Mechanics, Three-form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	221	11.550	.100	$.072<\sigma^2<.148$
Sources (s)	2	89.420	.020	$.011<\sigma^2<.060$
Forms (f)	2	152.005	.033	$.017<\sigma^2<.098$
Items within f (i:f)	15	15.846	.022	$.013<\sigma^2<.046$
ps	442	5.602	.278	$.246<\sigma^2<.316$
pf	442	1.172	.022	$.016<\sigma^2<.030$
sf	4	2.898	.001	$.001<\sigma^2<.004$
psf	884	.606	.053	$.045<\sigma^2<.062$
p(i:f)	3,315	.460	.057	$.050<\sigma^2<.065$
s(i:f)	30	.892	.003	$.002<\sigma^2<.006$
ps(i:f)	6,630	.290	.290	$.282<\sigma^2<.298$

Table A-6. Estimated Variance Components for Avionic Communication Specialists, Three-form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	49	7.910	.095	$.056<\sigma^2<.206$
Sources (s)	2	4.350	.001	$.001<\sigma^2<.001$
Forms (f)	2	29.790	.034	$.030<\sigma^2<.040$
Items within f (i:f)	12	3.986	.021	$.011<\sigma^2<.060$
ps	98	3.190	.179	$.060<\sigma^2<.243$
pf	98	.948	.016	$.008<\sigma^2<.028$
sf	4	.784	.000	$.000<\sigma^2<.000$
psf	196	.498	.041	$.029<\sigma^2<.066$
p(i:f)	588	.499	.069	$.055<\sigma^2<.091$
s(i:f)	24	.685	.008	$.004<\sigma^2<.022$
ps(i:f)	1,176	.291	.291	$.271<\sigma^2<.314$

Table A-7. Estimated Variance Components for Air Traffic Control Operators, Three-form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	179	6.300	.084	$.006<\sigma^2<.129$
Sources (s)	2	57.547	.025	$.013<\sigma^2<.074$
Forms (f)	2	47.170	.016	$.008<\sigma^2<.048$
Items within f (i:f)	9	12.390	.020	$.010<\sigma^2<.057$
ps	358	2.786	.189	$.163<\sigma^2<.222$
pf	358	.993	.028	$.019<\sigma^2<.044$
sf	4	.750	-.002	$.000<\sigma^2<.000$
psf	716	.516	.044	$.034<\sigma^2<.059$
p(i:f)	1,611	.483	.048	$.039<\sigma^2<.060$
s(i:f)	18	1.737	.008	$.005<\sigma^2<.017$
ps(i:f)	3,222	.339	.339	$.327<\sigma^2<.353$

Table A-8. Estimated Variance Components for Information Systems Radio Operators, Three-form Analysis

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	141	8.173	.106	$.078<\sigma^2<.156$
Sources (s)	2	2.569	-.001	$.000<\sigma^2<.000$
Forms (f)	2	48.237	.049	$.036<\sigma^2<.060$
Items within f (i:f)	12	8.535	.017	$.009<\sigma^2<.044$
ps	282	2.856	.157	$.133<\sigma^2<.187$
pf	282	1.033	.029	$.021<\sigma^2<.043$
sf	4	1.726	.001	$.000<\sigma^2<.002$
psf	564	.504	.040	$.032<\sigma^2<.053$
p(i:f)	1,692	.395	.031	$.024<\sigma^2<.041$
s(i:f)	24	1.103	.006	$.003<\sigma^2<.012$
ps(i:f)	3,384	.303	.303	$.291<\sigma^2<.315$

Table A-9. Simulated D Study Results of Ratings Analysis for Jet Engine Mechanics

σ^2 for ps(i:f) Design		σ^2 for pS(I:F) Design				
	n_r	1	1	1	3	3
	n_f	1	2	4	1	4
	n_i	8	4	8	12	8
$\sigma^2 p$=.100	$\sigma^2 p$=	.100	.100	.100	.100	.100
$\sigma^2 s$=.020	$\sigma^2 S$=	.020	.020	.020	.007	.007
$\sigma^2 f$=.033	$\sigma^2 F$=	.033	.017	.008	.033	.008
$\sigma^2 i:f$=.022	$\sigma^2 I:F$=	.003	.003	.001	.002	.001
$\sigma^2 s$=.278	$\sigma^2 pS$=	.278	.278	.278	.093	.093
$\sigma^2 pf$=.022	$\sigma^2 pF$=	.022	.011	.006	.022	.006
$\sigma^2 sf$=.001	$\sigma^2 SF$=	.001	.001	.000	.000	.000
$\sigma^2 psf$=.053	$\sigma^2 pSF$=	.053	.026	.013	.018	.004
$\sigma^2 p(i:f)$=.057	$\sigma^2 p(I:F)$=	.007	.007	.002	.005	.002
$\sigma^2 s(i:f)$=.003	$\sigma^2 S(I:F)$=	.000	.000	.000	.000	.000
$\sigma^2 ps(i:f)$=.290	$\sigma^2 pS(I:F)$=	.036	.036	.009	.009	.003
	σ^2 =	.100	.100	.100	.100	.100
	σ^2 =	.395	.358	.307	.147	.107
	σ^2 =	.453	.399	.336	.190	.123
	$\epsilon\rho^2$ =	.201	.217	.245	.405	.481
	Θ =	.180	.200	.228	.345	.447

Table A-10. Simulated D Study Results of Ratings Analysis for Avionic Communication Specialists

σ^2 for ps(i:f) Design		σ^2 for pS(I:F) Design				
	n_s	1	1	1	3	3
	n_f	1	2	4	1	4
	n_i	8	4	8	12	8
σ^2_p=.095	σ^2_p=	.095	.095	.095	.095	.095
σ^2_s=.001	σ^2_S=	.001	.001	.001	.000	.000
σ^2_f=.033	σ^2_F=	.034	.017	.009	.034	.008
$\sigma^2_{i:f}$=.021	$\sigma^2_{I:F}$=	.003	.003	.001	.002	.001
σ^2_{ps}=.179	σ^2_{pS}=	.179	.179	.179	.060	.060
σ^2_{pf}=.016	σ^2_{pF}=	.016	.008	.004	.016	.004
σ^2_{sf}=.000	σ^2_{SF}=	.000	.000	.000	.000	.000
σ^2_{psf}=.042	σ^2_{pSF}=	.042	.021	.010	.014	.003
$\sigma^2_{p(i:f)}$=.009	$\sigma^2_{p(I:F)}$=	.009	.009	.002	.001	.002
$\sigma^2_{s(i:f)}$=.008	$\sigma^2_{S(I:F)}$=	.001	.001	.002	.000	.000
$\sigma^2_{ps(i:f)}$=.291	$\sigma^2_{pS(I:F)}$=	.036	.036	.009	.009	.003
	σ^2 =	.095	.095	.095	.095	.095
	σ^2 =	.282	.253	.205	.100	.072
	σ^2 =	.320	.274	.215	.137	.082
	ϵP^2 =	.252	.273	.316	.487	.567
	Θ =	.228	.257	.305	.409	.536

Table A-11. Simulated D Study Results of Ratings Analysis for Air Traffic Control Operators

σ^2 for ps(i:f) Design		σ^2 for pS(I:F) Design				
	n_s	1	1	1	3	3
	n_f	1	2	4	1	4
	n_i	8	4	8	12	8
$\sigma^2 p$=.084	$\sigma^2 p$=	.084	.084	.084	.084	.084
$\sigma^2 s$=.025	$\sigma^2 S$=	.025	.025	.025	.008	.008
$\sigma^2 f$=.016	$\sigma^2 F$=	.016	.008	.004	.016	.004
$\sigma^2 i:f$=.019	$\sigma^2 I:F$=	.002	.002	.001	.002	.001
$\sigma^2 ps$=.189	$\sigma^2 pS$=	.189	.189	.189	.063	.063
$\sigma^2 pf$=.028	$\sigma^2 pF$=	.028	.014	.007	.028	.007
$\sigma^2 sf$=.000	$\sigma^2 SF$=	.000	.000	.000	.000	.000
$\sigma^2 psf$=.044	$\sigma^2 pSF$=	.044	.022	.011	.015	.004
$\sigma^2 p(i:f)$=.048	$\sigma^2 p(I:F)$=	.006	.006	.002	.004	.002
$\sigma^2 s(i:f)$=.008	$\sigma^2 S(I:F)$=	.001	.001	.000	.000	.000
$\sigma^2 ps(i:f)$=.340	$\sigma^2 pS(I:F)$=	.042	.042	.011	.009	.004
	σ^2 =	.084	.084	.084	.084	.084
	σ^2 =	.309	.274	.219	.119	.079
	σ^2 =	.353	.310	.249	.145	.092
	εP^2 =	.214	.236	.278	.414	.517
	Θ =	.193	.214	.253	.356	.479

Table A-12. Simulated D Study Results of Ratings Analysis for Information Systems Radio Operators

σ^2 for ps(i:f) Design		σ^2 for pS(I:F) Design				
	n_s	1	1	1	3	3
	n_f	1	2	4	1	4
	n_i	8	4	8	12	8
σ^2_p=.106	σ^2_p=	.106	.106	.106	.106	.106
σ^2_s=.000	σ^2_S=	.000	.000	.000	.000	.000
σ^2_f=.049	σ^2_F=	.049	.024	.012	.049	.012
$\sigma^2_{i:f}$=.017	$\sigma^2_{I:F}$=	.002	.002	.000	.001	.001
σ^2_{ps}=.157	σ^2_{pS}=	.157	.157	.157	.052	.052
σ^2_{pf}=.029	σ^2_{pF}=	.029	.015	.007	.029	.007
σ^2_{sf}=.001	σ^2_{SF}=	.001	.000	.000	.000	.000
σ^2_{psf}=.040	σ^2_{pSF}=	.040	.020	.010	.013	.003
$\sigma^2_{p(i:f)}$=.031	$\sigma^2_{p(I:F)}$=	.004	.004	.001	.003	.001
$\sigma^2_{s(i:f)}$=.006	$\sigma^2_{S(I:F)}$=	.001	.001	.000	.000	.000
$\sigma^2_{ps(i:f)}$=.303	$\sigma^2_{pS(I:F)}$=	.038	.038	.006	.009	.003
	σ^2 =	.106	.106	.106	.106	.106
	σ^2 =	.268	.233	.181	.107	.067
	σ^2 =	.320	.261	.194	.150	.080
	ϵP^2 =	.284	.313	.370	.498	.614
	Θ =	.249	.289	.354	.414	.571

Table A-13. Ġ Study Results for Crossed Design Analysis of WTPT Scores, Jet Engine Mechanics

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	254	.851	.008	$.006<\sigma^2<.011$
Method (m)	1	98.400	.012	$.006<\sigma^2<.034$
Tasks (t)	4	2.415	.000	$.000<\sigma^2<.000$
Items within t (i:t)	25	5.930	.000	$.000<\sigma^2<.000$
pm	254	.255	.002	$.001<\sigma^2<.005$
pt	1,016	.307	.008	$.006<\sigma^2<.011$
mt	4	9.685	.001	$.001<\sigma^2<.015$
pmt	1,016	.198	.012	$.010<\sigma^2<.015$
p(i:t)	6,350	.145	.009	$.007<\sigma^2<.013$
m(i:t)	25	7.562	.029	$.019<\sigma^2<.049$
pm(i:t)	6,350	.127	.127	$.123<\sigma^2<.131$

Table A-14. G Study Results for Crossed Design Analysis of WTPT Scores, Avionic Communication Specialists

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	95	1.172	.006	$.004<\sigma^2<.014$
Method (m)	1	50.704	.014	$.007<\sigma^2<.042$
Tasks (t)	5	23.652	.016	$.008<\sigma^2<.048$
Items within t (i:t)	30	4.907	.017	$.010<\sigma^2<.038$
pm	95	.355	.007	$.005<\sigma^2<.010$
pt	475	.486	.025	$.021<\sigma^2<.030$
mt	5	.989	-.001	$.000<\sigma^2<.000$
pmt	475	.119	.008	$.006<\sigma^2<.010$
p(i:t)	2,850	.139	.032	$.029<\sigma^2<.036$
m(i:t)	30	1.566	.016	$.011<\sigma^2<.025$
pm(i:t)	2,850	.074	.074	$.071<\sigma^2<.077$

Table A-15. G Study Results for Crossed Design Analysis of WTPT Scores, Air Traffic Control Operators

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	184	1.036	.007	$.005<\sigma^2<.012$
Method (m)	1	.151	.000	$.000<\sigma^2<.000$
Tasks (t)	5	20.535	.008	$.004<\sigma^2<.023$
Items within t (i:t)	24	5.616	.010	$.006<\sigma^2<.025$
pm	184	.113	.001	$.000<\sigma^2<.001$
pt	920	.580	.034	$.029<\sigma^2<.039$
mt	5	2.010	.000	$.000<\sigma^2<.001$
pmt	920	.098	.007	$.005<\sigma^2<.007$
p(i:t)	4,416	.211	.073	$.071<\sigma^2<.075$
m(i:t)	24	1.700	.009	$.006<\sigma^2<.015$
pm(i:t)	4,416	.065	.065	$.063<\sigma^2<.067$

Table A-16. G Study Results for Crossed Design Analysis of WTPT Scores, Information Systems Radio Operators

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	155	1.648	.032	$.025<\sigma^2<.042$
Method (m)	1	.878	.000	$.000<\sigma^2<.002$
Tasks (t)	4	11.475	.007	$.004<\sigma^2<.021$
Items within t (i:t)	15	1.887	.005	$.002<\sigma^2<.013$
pm	155	.128	.000	$.000<\sigma^2<.000$
pt	620	.378	.028	$.024<\sigma^2<.030$
mt	4	.622	.000	$.000<\sigma^2<.001$
pmt	620	.131	.020	$.017<\sigma^2<.023$
p(i:t)	2,325	.077	.012	$.010<\sigma^2<.015$
m(i:t)	15	.400	.002	$.001<\sigma^2<.005$
pm(i:t)	2,325	.052	.052	$.050<\sigma^2<.054$

Table A-17. G Study Results for Nested Design Analysis of WTPT Scores, Jet Engine Mechanics

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	254	1.248	.008	$.007<\sigma^2<.011$
Methods (m)	1	191.083	.013	$.007<\sigma^2<.038$
Tasks within m (t:m)	16	10.629	.003	$.002<\sigma^2<.010$
Items within t within m (i:t:m)	90	5.353	.020	$.016<\sigma^2<.027$
pm	254	.333	.001	$.001<\sigma^2<.003$
p(t:m)	4,064	.257	.019	$.017<\sigma^2<.021$
p(i:t:m)	22,860	.144	.144	$.141<\sigma^2<.146$

Table A-18. G Study Results for Nested Design Analysis of WTPT Scores, Avionic Communication Specialists

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	87	1.631	.013	$.010<\sigma^2<.018$
Methods (m)	1	16.200	.001	$.001<\sigma^2<.004$
Tasks within m (t:m)	20	8.949	.014	$.007<\sigma^2<.036$
Items within t within m (i:t:m)	88	2.780	.030	$.024<\sigma^2<.040$
pm	87	.207	-.001	$.000<\sigma^2<.000$
p(t:m)	1,740	.268	.032	$.029<\sigma^2<.035$
p(i:t:m)	7,656	.108	.108	$.105<\sigma^2<.110$

Table A-19. G Study Results for Nested Design Analysis of WTPT Scores, Air Traffic Control Operators

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	185	.775	.007	$.006<\sigma^2<.009$
Methods (m)	1	5.799	-.001	$.000<\sigma^2<.000$
Tasks within m (t:m)	12	18.821	.012	$.006<\sigma^2<.034$
Items within t within m (i:t:m)	70	5.741	.030	$.023<\sigma^2<.041$
pm	185	.164	-.002	$.000<\sigma^2<.000$
p(t:m)	2,220	.237	.018	$.016<\sigma^2<.020$
p(i:t:m)	12,950	.128	.128	$.126<\sigma^2<.131$

Table A-20. G Study Results for Nested Design Analysis of WTPT Scores, Information Systems Radio Operators

Effect	Df	Ms	σ^2	90% Confidence intervals
Persons (p)	156	2.704	.029	$.024<\sigma^2<.035$
Methods (m)	1	.057	-.001	$.000<\sigma^2<.000$
Tasks within m (t:m)	20	6.576	.008	$.004<\sigma^2<.018$
Items within t within m (i:t:m)	66	1.528	.009	$.007<\sigma^2<.013$
pm	156	.169	-.003	$.000<\sigma^2<.000$
p(t:m)	3,120	.283	.051	$.048<\sigma^2<.054$
p(i:t:m)	10,296	.080	.080	$.078<\sigma^2<.082$

Table A-21. Simulated D Study Results for WTPT Analysis of Jet Engine Mechanics

σ^2 for pm(i:t) Design		σ^2 for pM(I:T) Design				
	n_m	1	1	1	2	2
	n_t	5	10	10	5	15
	n_i	5	5	15	15	10
σ^2_p=.0081	σ^2_p=	.0081	.0081	.0081	.0081	.0081
σ^2_m=.0116	σ^2_M=	.0116	.0116	.0116	.0058	.0058
σ^2_t=.0000	σ^2_T=	.0000	.0000	.0000	.0000	.0000
$\sigma^2_{i:t}$=.0000	$\sigma^2_{I:T}$=	.0000	.0000	.0000	.0000	.0000
σ^2_{pm}=.0019	σ^2_{pM}=	.0019	.0019	.0019	.0010	.0010
σ^2_{pt}=.0075	σ^2_{pT}=	.0015	.0008	.0008	.0015	.0005
σ^2_{mt}=.0013	σ^2_{MT}=	.0003	.0001	.0001	.0001	.0000
σ^2_{pmt}=.0120	σ^2_{pMT}=	.0024	.0012	.0012	.0012	.0004
$\sigma^2_{p(i:t)}$=.0093	$\sigma^2_{p(I:T)}$=	.0004	.0002	.0001	.0001	.0001
$\sigma^2_{m(i:t)}$=.0292	$\sigma^2_{M(I:T)}$=	.0011	.0006	.0002	.0002	.0001
$\sigma^2_{pm(i:t)}$=.1267	$\sigma^2_{pM(I:T)}$=	.0050	.0025	.0008	.0008	.0004
	σ^2 =	.0081	.0081	.0081	.0081	.0081
	σ^2 =	.0112	.0066	.0048	.0046	.0023
	σ^2 =	.0243	.0189	.0167	.0107	.0083
	ϵP^2 =	.419	.553	.631	.637	.777
	Θ =	.252	.301	.327	.430	.495

Table A-22. Simulated D Study Results for WTPT Analysis of Avionic Communication Specialists

σ^2 for pm(i:t) Design		σ^2 for pM(I:T) Design				
	n_m	1	1	1	2	2
	n_t	5	10	10	5	15
	n_i	5	5	15	15	10
σ^2_p=.0062	σ^2_p=	.0062	.0062	.0062	.0062	.0062
σ^2_m=.0141	σ^2_M=	.0141	.0141	.0141	.0141	.0141
σ^2_t=.0160	σ^2_T=	.0032	.0016	.0016	.0032	.0011
$\sigma^2_{i:t}$=.0171	$\sigma^2_{I:T}$=	.0007	.0003	.0002	.0002	.0002
σ^2_{pm}=.0066	σ^2_{pM}=	.0066	.0066	.0066	.0033	.0033
σ^2_{pt}=.0252	σ^2_{pT}=	.0050	.0025	.0017	.0050	.0017
σ^2_{mt}=.0000	σ^2_{MT}=	.0000	.0000	.0000	.0000	.0000
σ^2_{pmt}=.0076	σ^2_{pMT}=	.0015	.0008	.0005	.0008	.0003
$\sigma^2_{p(i:t)}$=.0323	$\sigma^2_{p(I:T)}$=	.0013	.0007	.0004	.0004	.0004
$\sigma^2_{m(i:t)}$=.0155	$\sigma^2_{M(I:T)}$=	.0006	.0003	.0002	.0001	.0001
$\sigma^2_{pm(i:t)}$=.0740	$\sigma^2_{pM(I:T)}$=	.0030	.0015	.0010	.0005	.0005
	σ^2 =	.0062	.0062	.0062	.0062	.0062
	σ^2 =	.0174	.0120	.0102	.0100	.0061
	σ^2 =	.0360	.0283	.0258	.0205	.0146
	ϵP^2 =	.264	.343	.381	.384	.504
	Φ =	.148	.181	.195	.233	.300

Table A-23. Simulated D Study Results for WTPT Analysis of Air Traffic Control Operators

σ^2 for pm(i:t) Design		σ^2 for pM(I:T) Design				
	n_m	1	1	1	2	2
	n_t	5	10	10	5	15
	n_i	5	5	15	15	10
σ^2_p=.0073	σ^2_p=	.0073	.0073	.0073	.0073	.0073
σ^2_m=.0000	σ^2_M=	.0000	.0000	.0000	.0000	.0000
σ^2_t=.0077	σ^2_T=	.0015	.0008	.0008	.0015	.0005
$\sigma^2_{i:t}$=.0102	$\sigma^2_{I:T}$=	.0004	.0002	.0000	.0001	.0000
σ^2_{pm}=.0005	σ^2_{pM}=	.0005	.0005	.0005	.0003	.0003
σ^2_{pt}=.0336	σ^2_{pT}=	.0067	.0033	.0034	.0067	.0022
σ^2_{mt}=.0003	σ^2_{MT}=	.0000	.0000	.0000	.0000	.0000
σ^2_{pmt}=.0066	σ^2_{pMT}=	.0013	.0007	.0007	.0007	.0002
$\sigma^2_{p(i:t)}$=.0732	$\sigma^2_{p(I:T)}$=	.0029	.0015	.0005	.0010	.0005
$\sigma^2_{m(i:t)}$=.0089	$\sigma^2_{M(I:T)}$=	.0004	.0002	.0000	.0001	.0000
$\sigma^2_{pm(i:t)}$=.0649	$\sigma^2_{pM(I:T)}$=	.0026	.0013	.0004	.0004	.0002
	σ^2 =	.0073	.0073	.0073	.0073	.0073
	σ^2 =	.0141	.0073	.0054	.0090	.0034
	σ^2 =	.0164	.0085	.0064	.0108	.0040
	ϵP^2 =	.343	.502	.574	.449	.683
	Φ =	.309	.465	.535	.405	.645

Table A-24. Simulated D Study Results for WTPT Analysis of Information Systems Radio Operators

σ^2 for pm(i:t) Design		σ^2 for pM(I:T) Design				
	n_m	1	1	1	2	2
	n_t	5	10	10	5	15
	n_i	5	5	15	15	10
σ^2_p=.0318	σ^2_p=	.0318	.0318	.0318	.0318	.0318
σ^2_m=.0001	σ^2_M=	.0001	.0001	.0001	.0000	.0000
σ^2_t=.0073	σ^2_T=	.0015	.0007	.0007	.0015	.0005
$\sigma^2_{i:t}$=.0047	$\sigma^2_{I:T}$=	.0002	.0001	.0000	.0001	.0000
σ^2_{pm}=.0000	σ^2_{pM}=	.0000	.0000	.0000	.0000	.0000
σ^2_{pt}=.0278	σ^2_{pT}=	.0056	.0028	.0028	.0056	.0019
σ^2_{mt}=.0002	σ^2_{MT}=	.0002	.0002	.0002	.0000	.0000
σ^2_{pmt}=.0197	σ^2_{pMT}=	.0039	.0020	.0020	.0020	.0007
$\sigma^2_{p(i:t)}$=.0003	$\sigma^2_{p(I:T)}$=	.0005	.0005	.0001	.0002	.0001
$\sigma^2_{m(i:t)}$=.0022	$\sigma^2_{M(I:T)}$=	.0001	.0004	.0000	.0000	.0000
$\sigma^2_{pm(i:t)}$=.0519	$\sigma^2_{pM(I:T)}$=	.0021	.0010	.0004	.0004	.0002
	σ^2 =	.0318	.0318	.0318	.0318	.0318
	σ^2 =	.0121	.0060	.0052	.0080	.0028
	σ^2 =	.0140	.0070	.0061	.0096	.0033
	$\epsilon\rho^2$ =	.725	.841	.860	.798	.920
	θ =	.695	.819	.840	.767	.905